Proceedings of the Fourth International Workshop on Citizen-Centric Multiagent Systems 2026 (C-MAS 2026)

Co-located with the International Conference on Autonomous Agents and Multiagent Systems (AAMAS'26)

Paphos, Cyprus

Vahid Yazdanpanah[1], Nirav Ajmeri[2], Yali Du[3], Nadin Kokciyan[4], Fernando P. Santos[5], and Sebastian Stein[1]

[1]University of Southampton
[2]University of Bristol
[3]King's College London
[4]University of Edinburgh
[5]University of Amsterdam

26 May 2026

Welcome to the fourth edition of C-MAS, the International Workshop on Citizen-Centric Multiagent Systems. C-MAS continues to explore how multiagent systems, autonomous agents, and AI-based sociotechnical systems can be designed around citizens as active participants rather than passive users, data sources, or service recipients. As AI systems increasingly mediate access to public services, information, mobility, finance, healthcare, and collective decision-making, it becomes essential to understand how citizens' preferences, values, rights, vulnerabilities, and strategic behaviours can be represented and respected.

C-MAS 2026 builds on the foundations established in previous editions by broadening the discussion around citizen agency, accountability, fairness, and participation. This year's accepted papers address a diverse set of topics, including accountability and explainability in citizen-centric MAS, emotionally intelligent human-AI interaction, biased social norms in large language models, long-term fairness dynamics, strategic behaviour in school choice and lending, multiagent reinforcement learning for cooperation and logistics, human-centric mobility services, algorithmic influence in information diffusion, and the realism of generative agents in social simulations.

The workshop also features a keynote by Dr. Roxana Rădulescu on human-aligned agents and multi-objective reinforcement learning. The keynote highlights a central

challenge for citizen-centric AI, namely that many socially relevant problems involve multiple stakeholders, conflicting objectives, and trade-offs that cannot be reduced to a single reward signal. This perspective strongly resonates with the themes of C-MAS 2026, where the design of AI and multiagent systems requires not only technical performance or the optimisation of a single aspect, but also attention to transparency, trust, fairness, and human values.

We hope these proceedings provide a useful snapshot of current research on citizen-centric multiagent systems and help foster further collaboration across AI, multiagent systems, social simulation, responsible AI, public policy, and human-centred design. We thank all authors, reviewers, organisers, session chairs, and participants for contributing to the continuing development of the C-MAS community.

Further details about C-MAS 2026 are available on the workshop webpage:
https://sites.google.com/view/cmas2026

Contents

1 Keynote: From Friction to Synergy: Building Human-Aligned Agents with Multi-Objective Reinforcement Learning **4**

2 Norms, Accountability, and Human-Aligned Agents **5**
2.1 Explainability through Accountability in Citizen-Centric Multiagent Systems 5
2.2 A Citizen-Centric Multi-Agent Framework for Emotionally Intelligent Human–AI Interaction 13
2.3 Biased Social Norms of Cooperation in Large Language Models . . . 21

3 Fairness, Institutions, and Strategic Public Decision-Making **29**
3.1 Long-term Effects of Fairness Metrics on Population Dynamics 29
3.2 Strategic Exploitation of Placement Guarantees in Random Serial Dictatorship: Modelling the Amsterdam School Choice System 39
3.3 The Role of Regulatory Institutions in Strategic Lending 49

4 Multiagent Learning, Cooperation, and Coordination **57**
4.1 Resolving Complex Social Dilemmas by Aligning Preferences with Counterfactual Regret 57
4.2 Tripartite Tender Game Framework for MARL in Crowdsource First-Last Mile Logistics 84
4.3 Human-Centric AI Modeling for Fair and Cooperative Microtransit Ride Prioritization 101

5 Generative and Algorithmic Influence in Sociotechnical Systems **109**
5.1 Algorithmic Recommendations Alter the Spread of Competing Information 109
5.2 The Challenge of Realistic Personas in Generative Agent-Based Modeling 125

1 Keynote: From Friction to Synergy: Building Human-Aligned Agents with Multi-Objective Reinforcement Learning

Keynote by Dr. Roxana Rădulescu, Utrecht University

Most complex problems of social relevance, such as climate change mitigation, taxation policy design, and traffic management, are inherently multi-agent and multi-objective, involving diverse stakeholders with frequently conflicting goals. Building human-aligned agents in these domains requires a framework capable of navigating the inevitable friction between disparate objectives. While Reinforcement Learning has become a pivotal tool for designing solutions in these critical areas, traditional RL often falls short by collapsing complex trade-offs into a single scalar reward. In this talk, Dr. Roxana Rădulescu discusses how Multi-Objective Reinforcement Learning (MORL) offers a more robust and adaptable alternative by explicitly modeling the multi-dimensional nature of feedback signals. She presents MORL as a foundational framework for conflict-aware systems, showcasing how it can foster key principles like explainability, transparency, and trust. Finally, she explores how this multi-objective approach provides a flexible mechanism for humans and agents to co-evolve solutions, turning systemic conflict into collaborative synergy.

Dr. Roxana Rădulescu is an Assistant Professor in AI and Data Science at the Department of Information and Computing Sciences at Utrecht University. Her research focuses on reinforcement learning and multi-agent systems, with particular emphasis on multi-objective decision-making and multi-objective multi-agent reinforcement learning (MOMARL), where autonomous agents must balance multiple, often conflicting objectives. She received her PhD in Computer Science (Artificial Intelligence) from Vrije Universiteit Brussel, where her work developed a utility-based perspective on decision-making in multi-objective multi-agent systems. Prior to joining Utrecht University, she was a postdoctoral researcher at the Artificial Intelligence Lab at Vrije Universiteit Brussel supported by an FWO fellowship. Her work spans reinforcement learning, game theory, and multi-objective optimisation, and has been published in leading venues such as JAAMAS and AAMAS. She is also actively involved in the AI research community through tutorials at major conferences and service roles including organizing committees for AAMAS, IJCAI, and ECAI.

2 Norms, Accountability, and Human-Aligned Agents

2.1 Explainability through Accountability in Citizen-Centric Multiagent Systems

Explainability through Accountability in Citizen-Centric Multiagent Systems

Matteo Baldoni[1 [0000−0002−9294−0408]], Cristina Baroglio[1 [0000−0002−2070−0616]], Elisa Marengo[1 [0000−0003−1879−2088]], Roberto Micalizio[1 [0000−0001−9336−0651]], and Stefano Tedeschi[2 [0000−0002−9861−390X]]

[1] Università degli Studi di Torino - Dipartimento di Informatica, Torino, Italy
{matteo.baldoni,cristina.baroglio,elisa.marengo,roberto.micalizio}@unito.it
[2] Università della Valle d'Aosta - Université de la Vallée d'Aoste, Aosta, Italy
s.tedeschi@univda.it

Abstract. AI is increasingly used in critical areas like healthcare, energy, and disaster response, where trust and transparency are essential. Current explainable AI approaches treat citizens as passive users, ignoring their stakeholder roles and expectations. This paper argues for accountability as a core principle in multi-agent systems (MAS), going beyond *post hoc* justification of decisions. Accountability creates a socio-technical framework for continuous feedback and decision interpretation among all agents –human and AI. It supports meaningful explanations, fairer outcomes, and participatory processes, empowering citizens as active participants. By integrating organizational models, accountability ensure AI systems align with societal needs and values.

Keywords: Accountability · Explanation · Organizations.

1 Introduction

Large-scale AI systems are increasingly embedded in socio-technical contexts that shape citizens' everyday lives, such as energy management, urban mobility, or emergency response (see, e.g., [1, 22]). Here, AI must not only be effective, but also trustworthy, fair, and aligned with human values. Yet, citizens are often treated as passive data providers or service consumers rather than as stakeholders in system design. Multi-agent Systems (MAS) [29] represent a step forward towards encompassing citizens in a value-respectful way because they explicitly represent autonomous and interacting entities (the agents), that may well amount to citizens, service providers, infrastructures, and institutions. Despite having proved effective in software engineering and business process modeling, the autonomy and distribution characteristic of MAS still pose challenges to transparency and responsibility, especially when decisions have social consequences, calling for further means that involve explanations and accountability.

Explainable AI (xAI) [24] has emerged to address these issues by providing human-understandable explanations of system behavior. Indeed, AI systems, especially those based on machine learning, are often opaque: sometimes even

the designers cannot determine *why* the system came up with a decision in high-level, meaningful terms, rather than relying on mathematical considerations. This opacity extends to MAS, where agents are, by their own nature, autonomous. In this context, the need of providing agents and MAS with capabilities and infrastructures that make decisions/behaviors explainable, emerges not only to provide the end user with human-understandable explanations, but also from a software engineering perspective because the exchange of explanations between components is functional to the system's overall objectives. Yet, in complex MAS, explanations alone are insufficient because they are context-dependent, audience-specific, and often disconnected from the organizational structures – that does not determine *who should provide explanations, to whom, and for what purpose*. The need for explanation is, indeed, part of a broader problem, related to the absence of properly devised channels for collecting and propagating feedback about agents' decisions/actions through a network of autonomous, yet interconnected, parts.

In this paper, we advance two main arguments. First, we contend that accountability is a fundamental requirement for *citizen-centric multi-agent systems* (C-MAS), as it provides the socio-technical infrastructure through which explanations can be generated, situated, contested, and leveraged for learning and adaptation. Second, we argue for a shift from a sole focus on explainability toward the explicit modeling of accountability relationships among agents, with citizens treated as first-class stakeholders. These perspectives are complementary but not equivalent. We posit that accountability should not be understood merely as a means to enhance explainability; rather, it constitutes the underlying infrastructure that enables explanations to emerge, evolve, and be meaningfully used. In this sense, explainability is better conceived as an outcome, whereas accountability provides the mechanisms and structures through which such outcomes can be systematically realized.

2 Explanation and Explainability

Explanation has long been studied in philosophy, psychology, and cognitive sciences [27]. It can be analyzed at two complementary levels: the *process of explaining*, and the resulting *explanatory product*. Explaining is commonly understood as providing an answer to a "why" question about an event or decision [21]. In C-MAS, such a need naturally arises when citizens and other stakeholders seek to understand system behavior. In AI, this view dates back to Reiter's work on Model-Based Diagnosis (MBD) [26], where explanations are derived by interpreting observed behavior (i.e., symptoms) with respect to a system model. The expressive power of the model shapes the explanations that can be produced: models of normal behavior yield consistency-based explanations, while models including abnormal behavior support abductive explanations that capture underlying causes [14]. While these approaches provide a solid foundation for reasoning about explanations, their application to C-MAS is complex because of the heterogeneity of stakeholders, that include citizens, service providers, and

public authorities, each with their roles, knowledge, and legitimate expectations. Moreover, agents typically have only partial views of the overall system, which limits their individual ability to produce explanations.

Structurally, explanations distinguish an *explanandum*, the phenomenon to be explained, and an *explanans*, i.e., the set of statements adduced to make sense of the explanandum. A *scientific* explanation is considered sound when: (i) the explanandum logically follows from the explanans; (ii) the explanans includes general laws; (iii) the explanans has empirical content; and (iv) the statements constituting the explanans are true [21]. However, some scholars argued that formal soundness alone is insufficient if explanations fail to address stakeholders' informational needs: following [27], explanation is inherently audience- and context-dependent, and what is included, omitted, or taken for granted depends on the intended recipients and their background knowledge. As a result, explanations are often partial and selective, even when they are formally correct. This aspect is particularly relevant in C-MAS, where citizens may lack technical expertise but remain entitled to understand. This pragmatic view is further articulated by [23], who characterizes explanation as involving both a *cognitive process*, identifying potentially relevant causes of an event (often via counterfactual reasoning), and a *social process* of knowledge transfer between explainer and explainee. This social dimension is particularly salient in C-MAS, where explanations must support coordination, trust, and informed decision-making.

2.1 The Limits of the Approach with an Example

In distributed and autonomous settings such as C-MAS, actions and decisions depend on the availability of contextual information that is often fragmented across agents. Thus, the causal dependencies required for meaningful explanations may be unavailable to the agent that detected a failure, since each agent has a partial view of the system and of the social implications of its decisions.

Consider an on-demand mobility service designed to complement public transportation in underserved areas. The system coordinates a fleet of shared vehicles through a MAS architecture involving *citizen agents* (representing individual travelers), *dispatch agents* (operated by the service provider), and *policy agents* (representing constraints imposed by local authorities, such as equity or environmental goals). Citizens submit ride requests specifying pickup and drop-off locations, time constraints, and optionally accessibility requirements. During peak hours, a citizen's ride request may not be assigned within the expected time window. A naïve explanation – *no vehicle was available* – may be technically correct, yet insufficient in a citizen-centric context. The lack of assignment may result from multiple, context-dependent factors: the citizen may have specified accessibility constraints that only a subset of vehicles can satisfy; policy agents may have temporarily prioritized vulnerable users or essential workers; or real-time traffic conditions may have invalidated previously feasible routes. None of these factors is necessarily visible to the dispatch agent in isolation, nor are they apparent to the citizen receiving the explanation. If the dispatch agent is asked to explain the failure, it can only report information available within

its local context, such as the absence of feasible vehicle assignments under its current optimization model. Although correct, this explanation does not enable the citizen to understand whether the outcome was due to personal constraints, broader policy decisions, or system limitations. Nor does it support the system in determining whether the failure reveals a need for improved preference modeling, revised incentive mechanisms, or policy adjustments.

This example illustrates that explanations produced by individual agents are rarely sufficient on their own. Their usefulness depends on the availability of structured mechanisms for gathering, propagating, and interpreting context-rich information across agents with different roles and responsibilities. In the following, we argue that accountability, and in particular the interpretation in [2, 4–7, 9], provides the necessary infrastructure for enabling such mechanisms.

3 Accountability as a Foundation for C-MAS

Accountability plays a central role in human organizations, yet it is a multi-faceted concept with multiple interpretations. In this work, we adopt a notion of accountability that supports explanation, learning, adaptation, and innovation – capabilities that are all essential for C-MAS. Our perspective is inspired by organizational frameworks such as UNDP [19], where accountability is realized through processes for monitoring, evaluating, and improving performance based on evidence. These frameworks define roles, responsibilities, and procedures to foster continuous learning and capacity-building, treating deviations and failures as opportunities for reflection and improvement. Accountability also enables oversight by making information available for assessing compliance with regulatory and ethical standards. Despite domain differences, similar principles emerge across organizations (e.g., [25, 28]). Here, accountability provides the infrastructure through which relevant information is collected, shared, and interpreted by authorized actors, supporting a feedback loop that links actions to reporting and learning. This understanding is established in sociology where accountability is a key feature of governance grounded in shared expectations [18], and extends beyond blame to include the capacity and legitimacy to provide accounts [17], becoming a mechanism for the constitution of social order [20].

Computationally, accountability can be modeled as a relationship between an account taker (*a-taker*), who is entitled to request an account, and an account giver (*a-giver*), who is obliged to provide it [3, 9]. In C-MAS, it offers a principled way to represent citizens, service providers, and institutions as stakeholders with legitimate claims and responsibilities. While related to explainability, accountability is not equivalent to it: accounts are structured, context-rich information about actions and conditions, while explanations are interpretive constructs tailored to specific audiences. Accountability ensures that accounts are available, traceable, and accessible, enabling the construction of explanations when needed.

Returning to the mobility scenario, accountability relations can be established along the organizational structure of the system. For instance, vehicle agents are accountable to dispatch agents with respect to their status (e.g.,

availability, faults, delays); policy agents are accountable to dispatch agents regarding the activation of prioritization rules; and dispatch agents are accountable to citizen agents for allocation decisions. When a ride request is not fulfilled, the citizen agent (a-taker) can request an account from the dispatch agent (a-giver). In turn, the dispatch agent may request accounts from other agents to reconstruct the relevant context. Through this chain of accountability, the dispatch agent can assemble a structured, context-rich account of the event. For example, it may combine information about vehicle unavailability, policy-driven prioritization, and traffic disruptions into a coherent account explaining why the request could not be satisfied within the expected time window. Each contributing agent provides information for which it is accountable, ensuring traceability and reliability. The resulting explanation is no longer a simplistic statement, but a grounded and interpretable reconstruction of the decision context. E.g., the system may report that the request could not be fulfilled because only accessible vehicles were suitable, one was delayed due to a fault, and the remaining ones were allocated to higher-priority users according to active policies. Since the citizen agent is modeled as an a-taker with appropriate interpretive capabilities, such an account can be tailored to be both informative and understandable.

More generally, making accountability relationships explicit enables the system to systematically gather and propagate the information needed to explain complex outcomes. It also supports adaptive behavior: if recurring patterns of unmet requests are identified through accounts, dispatch or policy agents may revise allocation strategies or constraints. Thus, accountability not only enhances explainability but provides the socio-technical infrastructure through which explanations can be constructed, validated, and used for continuous improvement.

Accountability has both a *normative dimension*, capturing the legitimacy of requesting and providing accounts, and a *structural dimension*, capturing the control required to produce them. Control implies that agents can provide accounts either because they are directly involved in an event or because they can obtain relevant information from others [8, 9]. This view is supported by the information model in [6], which specifies the data needed to identify accountable agents. These characteristics make accountability particularly well suited to MAS. The inherently distributed nature of MAS, where knowledge, responsibilities, and decision-making are fragmented across agents, naturally aligns with accountability relationships, which organize how information is requested, produced, and interpreted among agents with different roles. In this sense, accountability provides a principled mechanism to connect partial views into context-rich accounts. Specifically, multi-agent organizations (MAOs) [10, 15, 16] offer a natural abstraction to operationalize this integration, structuring interactions through roles, norms, and communication channels. While norms ensure coordination, they are insufficient in dynamic, citizen-centric contexts where adaptation is required. By complementing norms, accountability explicitly structures how information flows across agents, enabling context-aware decisions, escalation, and adaptation. Importantly, this integration turns MAS into a powerful foundation for xAI. By ensuring that relevant information is traceable, accessi-

ble, and interpretable through accountability relationships, the system can systematically construct explanations grounded in distributed evidence. Through monitoring, auditing, and continuous improvement, accountability allows also organizational evolution in response to changing conditions and citizens' needs.

4 Conclusions

This paper presented accountability as a foundational concept for designing citizen-centric sociotechnical systems, in particular C-MAS. Accountability provides the infrastructure that gives explanations meaning and impact, enhancing fairness and legitimacy by tracking decision rationales and deviations across agents and organizational layers. It enables participatory feedback loops through which stakholders can shape system evolution and adapt to societal change.

Accountability is widely recognized as a cornerstone of effective human organizations, including international institutions, and has recently gained attention in software engineering and MAS design [11, 12]. An additional, often overlooked contribution is its role in innovation. Innovation frequently emerges from deviations from prescribed norms [13], which expose mismatches between assumptions and real-world conditions; through accountability, such deviations can be interpreted to revise objectives and organizational structures.

In previous work, we established the foundations for treating accountability as a first-class concept in agent organizations, demonstrating its contribution to robustness under known perturbations [2, 6, 8, 9]. We developed conceptual models, organizational abstractions, and programming patterns that operationalize accountability by linking specifications to agent behavior. Our results show that accountability complements normative coordination, allowing MAS not only to enforce expected behavior but also to adapt to unexpected situations and citizen input. As AI systems increasingly mediate interactions between institutions and citizens, integrating accountability into their design is essential to achieve explainability, trust, and continuous innovation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bajwa, A.: Ai-based emergency response systems: A systematic literature review on smart infrastructure safety. Available at SSRN 5171521 (2025)
2. Baldoni, M., Baroglio, C., Micalizio, R., Tedeschi, S.: Robustness based on Accountability in Multiagent Organizations. In: Proc. of AAMAS 2021 (2021)
3. Baldoni, M., Baroglio, C., Boissier, O., May, K.M., Micalizio, R., Tedeschi, S.: Accountability and Responsibility in Agents Organizations. In: Proc. of PRIMA 2018. No. 11224 in LNCS (2018)
4. Baldoni, M., Baroglio, C., Boissier, O., Micalizio, R., Tedeschi, S.: Accountability and responsibility in multiagent organizations for engineering business processes. In: Post-Proc. of EMAS@AAMAS 2019 (2020)

5. Baldoni, M., Baroglio, C., May, K.M., Micalizio, R., Tedeschi, S.: Computational accountability. In: Proc. of URANIA@AIxIA 2016. vol. 1802. CEUR-WS.org (2016)
6. Baldoni, M., Baroglio, C., May, K.M., Micalizio, R., Tedeschi, S.: MOCA: An ORM MOdel for Computational Accountability. Intell. Artif. **13**(1) (2019)
7. Baldoni, M., Baroglio, C., Micalizio, R.: Fragility and Robustness in Multiagent Systems. In: Post-Proc. of EMAS@AAMAS 2020. LNAI (2020)
8. Baldoni, M., Baroglio, C., Micalizio, R., Tedeschi, S.: Reimagining Robust Distributed Systems through Accountable MAS. IEEE Int. Comp. **25**(6) (2021)
9. Baldoni, M., Baroglio, C., Micalizio, R., Tedeschi, S.: Accountability in Multi-Agent Organizations: from Conceptual Design to Agent Programming. Autonomous Agents and Multi-Agent Systems **37**(7) (2023)
10. Boissier, O., Bordini, R.H., Hübner, J.F., Ricci, A., Santi, A.: Multi-agent oriented programming with JaCaMo. Sci. of Comp. Prog. **78**(6) (2013)
11. Chopra, A.K., Singh, M.P.: The thing itself speaks: Accountability as a foundation for requirements in sociotechnical systems. In: Proc. of RELAW 2014) (2014)
12. Chopra, A.K., Singh, M.P.: From social machines to social protocols: Software engineering foundations for sociotechnical systems. In: Proc. of WWW 2016 (2016)
13. Chopra, A.K., Singh, M.P.: Sociotechnical Systems and Ethics in the Large. In: Proc. of AIES 2018 (2018)
14. Console, L., Torasso, P.: A spectrum of logical definitions of model-based diagnosis. Computational Intelligence **7**(3) (1991)
15. Dastani, M., Tinnemeier, N.A., Meyer, J.J.C.: A programming language for normative multi-agent systems. In: Handbook of Research on Multi-Agent Systems: semantics and dynamics of organizational models (2009)
16. Dignum, V.: Handbook of Research on Multi-agent Systems: Semantics and Dynamics of Organizational Models (2009)
17. Dubnick, M.J.: Blameworthiness, trustworthiness, and the second-personal standpoint: Foundations for an ethical theory of accountability (2013)
18. Dubnick, M.J., Justice, J.B.: Accounting for accountability (2004)
19. Executive Board of the United Nations Development Programme and of the United Nations Population Fund: The UNDP accountability system, accountability framework and oversight policy. Tech. Rep. DP/2008/16/Rev.1, United Nations (2008)
20. Garfinkel, H.: Studies in ethnomethodology (1967)
21. Hempel, C.G., Oppenheim, P.: Studies in the logic of explanation. Philosophy of science **15**(2) (1948)
22. Javed, H., Eid, F., El-Sappagh, S., Abuhmed, T.: Sustainable energy management in the AI era: a comprehensive analysis of ML and DL approaches. Computing **107**(6) (2025)
23. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267** (2019)
24. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review **55**(5) (2022)
25. Office of the Auditor General of Canada: 2002 December Report of the Auditor General of Canada: Chapter 9 (2002)
26. Reiter, R.: A theory of diagnosis from first principles. Art. Int. **32**(1) (1987)
27. Ruben, D.H.: Explaining explanation (2015)
28. United Nations Children's Fund: Report on the accountability system of UNICEF (2009), e/ICEF/2009/15
29. Wooldridge, M.J.: Introduction to multiagent systems (2002)

2.2 A Citizen-Centric Multi-Agent Framework for Emotionally Intelligent Human–AI Interaction

A Citizen-Centric Multi-Agent Framework for Emotionally Intelligent Human–AI Interaction

Sam Nallaperuma-Herzberg[1], Rishabh Balse[1], Sonia Koszut[1], Lilith Stenhouse[1], Anna Bevan[1], Tristan Bekinschtein[1], and Pietro Lio[1]

University of Cambridge, UK

Abstract. Citizen-facing AI systems are increasingly being deployed in healthcare, education, and public services, where users are active stakeholders with emotional states, preferences, and vulnerabilities. Yet most conversational AI systems remain monolithic pipelines that optimize linguistic fluency while offering limited support for affect perception, adaptive deliberation, and normative self-governance. We propose a citizen-centric multi-agent framework that operationalizes emotional intelligence through four interacting agents: (i) a Perception agent that infers user affect from physiological and conversational cues, (ii) a Planner agent that selects interaction strategies via reinforcement learning, (iii) a Generator agent that produces strategy-conditioned responses, and (iv) a Metacognitive Critic agent that enforces safety and empathy norms before release. We instantiate the framework in digital mental health support and evaluate it through a blinded human study using the CARE empathy measure. Results indicate that treating citizens' affective states as first-class elements in a norm-aware multi-agent system improves perceived empathy, trust, and interaction quality.

Keywords: Citizen-centric AI · multi-agent systems · affective computing · reinforcement learning · trustworthy AI

1 Introduction

AI systems increasingly mediate interactions between citizens and essential services such as healthcare, education, and mental well-being support. In these settings, users are not passive recipients of automated outputs but active participants whose emotional states and expectations shape interaction outcomes. Citizen-centric multi-agent systems (MAS) explicitly emphasise modelling and reasoning about such human stakeholders to ensure trustworthy, fair, and socially responsible AI.

Despite advances in large language models, most conversational AI systems remain architecturally monolithic, focusing on surface-level fluency while lacking mechanisms to perceive user affect, adapt interaction strategies, or regulate outputs according to social norms. This limitation is especially problematic in digital mental health, where empathy, safety, and trust are critical.

From a MAS perspective, these limitations are structural. Classical work on Belief–Desire–Intention (BDI) agents shows that socially competent behavior

arises from agents with explicit mental states and deliberative control [13]. Normative MAS research further emphasizes the role of norms and governance in sociotechnical systems [1]. Treating users as passive endpoints rather than as agents with evolving affective states undermines these principles.

We propose a citizen-centric multi-agent architecture that closes an affective–cognitive loop—*human* →*perceiving agent* → *planning agent* → *generating agent* → *critique agent* → *human*—and embeds citizen affect and normative oversight directly into the system's decision-making process.

2 Preliminaries

Two key aspects of human emotional intelligence are empathy and meta-cognition. Empathy is the ability to understand the mental state of the others which help humans to adapt to others' state of mind. Meta-cognition considers monitoring and controlling of our own thoughts and behaviour [5].

Empathy in AI Agents Pure text sentiment misses activation and stress markers. State of the art EEG affect work reports θ, α and β waves corresponding to stress [8]. Combining these with linguistic cues yields a richer affective state space.

Meta-cognition in AI Agents In the context of digital therapy meta-cognition can correspond to filtering inaccurate or unsafe therapeutic responses. Prior digital therapy work relies on static toxicity filters. Meanwhile, critic guided approaches [9, 6] show effectiveness in automatically flagging hallucinated or off-topic replies.

Measuring the Effectiveness of Emotionally Intelligent Agents The Consultation and Relational Empathy (CARE) measure was developed by Mercer et al. (2004) [12] to assess perceptions of relational empathy in consultations. It consists of 10 items, each rated on a 5-point Likert scale from "Poor" to "Excellent." The measure evaluates aspects such as making the patient feel at ease, really listening, and showing care and compassion. The CARE Measure has demonstrated high internal reliability (Cronbach's alpha = 0.92) and strong validity across diverse patient populations .

No prior digital therapy system unifies neurofeedback powered empathy, meta-cognitive self-repair and LLM fluency. The proposed MAS system fills this gap.

3 Citizen-Centric Multi-Agent Architecture

The framework consists of four cooperating agents that communicate via structured message passing over a shared affective belief base. At each dialogue turn t, the system maintains an affective state

$$S_t = \langle Stress_t, Sent_t \rangle, \tag{1}$$

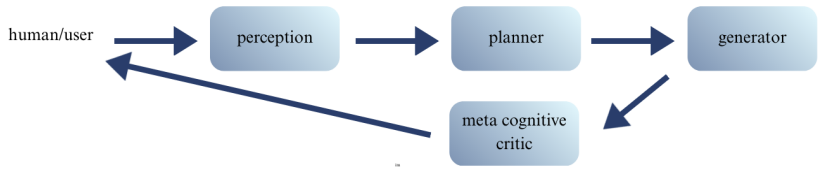


Fig. 1. High level overview of the proposed MAS.

representing the citizen's estimated stress and sentiment. The Perception agent updates S_t, the Planner selects a strategy, the Generator proposes a response, and the Critic evaluates it before release (see Figure 1).

Perception Agent The Perception agent infers affect using multimodal evidence. Physiological stress is estimated from multichannel EEG using a spatial–temporal neural model that combines temporal and graph convolutions. EEG signals sampled at 128 Hz are processed in sliding windows (5s length, 2s overlap, updated every 7s). For each window, the voltage matrix $V^{(i)} \in R^{C\times T_s}$ is passed to a Spatial–Temporal Residual Stress network,

$$s^{(i)} = f\big(V^{(i)}\big),$$

which outputs a continuous stress estimate $Stress_t$ for real-time monitoring.

Emotional tone in text is inferred using a BERT-based classifier [3] trained on the GoEmotions dataset [2]. Emotion class probabilities P are aggregated into a scalar sentiment score:

$$Sent_t = \tfrac{1}{2}\big(P_{\text{optimism}} + P_{\text{joy}} - P_{\text{sad}} - P_{\text{angry}} + 1\big). \tag{2}$$

Planner Agent The Planner agent adapts interaction strategies based on affect. It is formulated as a reinforcement learning agent [15], where the state is S_t and actions correspond to discrete conversational stances (e.g., supportive, neutral, challenging). The reward reflects citizen well-being:

$$R_t = -\Delta Stress_t + \Delta Sent_t. \tag{3}$$

We employ proximal policy optimisation (PPO) algorithm [14] to find the best policy that maximises the cumulative future reward. The policy is trained on behavioural cloning (BC) on a dataset of 5,000 transitions, minimising cross-entropy between expert and policy actions, and subsequently fine-tuned using PPO for 200 epochs.

Generator Agent The Generator agent is responsible for *realising* the planner's selected strategy as natural language. It conditions generation on (i) the dialogue context, (ii) the current affective state S_t, and (iii) the planner-selected strategy. Concretely, strategy is injected as a short system-level instruction so that the generator can modulate tone and intervention style without changing the user's surface input. This separation supports citizen-centric adaptation: the same user message can yield different helpful responses depending on stress/sentiment and

the chosen stance (see Table 1). The generator LLM is based on Llama-3.1-8B and fine-tuned using GaLore [17] on the CACTUS dataset [10]. The CACTUS dataset is a multi-turn dialogue corpus that emulates real-world interactions between a counselor and a client using Cognitive Behavioral Therapy (CBT).

Input	`If I don't find any friends, my life will be over.`
Base Model	*"I'm so sorry to hear you're feeling that way. It's understandable to feel lonely or isolated at times, and it can be really tough when we don't have people in our lives who understand us. But here's the thing: your life is not defined by whether or not you have friends. You are so much more than your social connections..."* (continues with lengthy generic reassurance)
Fine-tuned Model	*"It sounds like you're feeling quite overwhelmed by this thought. Can you tell me more about what's making you think that not having friends means your life is over?"*

Table 1. Example illustrating the difference in response styles between the Base and Fine-tuned Model.

Metacognitive Critic Agent The Metacognitive Critic provides *normative oversight* before responses reach the citizen. It evaluates the Generator's output for safety, coherence, and empathetic alignment, requesting revisions when necessary. This operationalizes normative MAS principles, where behavior is constrained by explicit governance mechanisms [4]. The critique loop is bounded to ensure liveness: the system either returns an approved response or falls back to a safe-harbour message. Toxicity is assessed using a RoBERTa classifier trained on the Jigsaw dataset [11, 7]. The design is compatible with agentic LLM frameworks that support iterative reasoning and refinement [16].

4 Application and Human Study

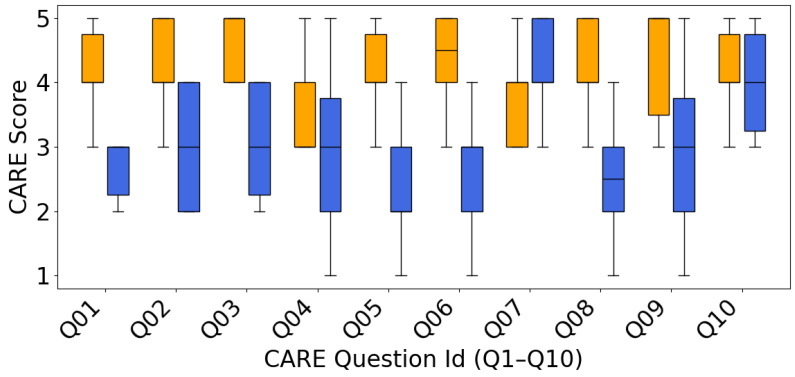


Fig. 2. Boxplots showing the comparative CARE scores for the proposed system (orange) versus baseline system (blue) rated by human participants

We instantiate the framework in digital mental health support, a citizen-centric domain where emotional intelligence and trust are essential. In a blinded within-subject study, 30 participants interacted with the proposed system and a baseline conversational system for 20 minutes each.

Perceived empathy was assessed using the CARE measure [12]. Table 2 presents the questions based on the CARE metric that are considered in this study.

Id	Measure
1	Making you feel at ease (welcoming, non-judgemental opening)
2	Letting you tell your story (allowing you to explain in your own words)
3	Really listening (showing interest in what you say and how you feel)
4	Being interested in you as a whole person (acknowledging broader context)
5	Fully understanding your concerns (grasping both content and affect)
6	Showing care and compassion (supportive without forcing solutions)
7	Being positive (encouraging hope without superficial reassurance)
8	Explaining things clearly (making patterns/beliefs understandable)
9	Helping you take control (supporting user-generated steps)
10	Making a plan of action with you (collaborative next steps)

Table 2. Adapted CARE measure items used in this study [12].

Figure 2 presents the results of the study. Participants consistently rated the proposed system as more empathetic. The median CARE score increased from 30.0 (baseline) to 42.0 (proposed), yielding a median paired improvement of 11.5 points. An exact paired Wilcoxon signed-rank test showed statistical significance ($p = 0.002$), with all participants preferring the proposed system. Qualitative feedback highlighted greater emotional awareness, attentiveness, and human-likeness.

5 Conclusion

We presented a citizen-centric multi-agent framework for emotionally intelligent human–AI interaction that integrates affective perception, reinforcement-learning-based planning, strategy-conditioned language generation, and a separate metacognitive critic for normative oversight. By treating citizens' affective states as first-class elements within a MAS architecture, the system improves perceived empathy and trust in a digital mental health setting. This work demonstrates how core MAS principles can support the design of trustworthy, human-centered AI systems.

Bibliography

[1] g. andrighetto, G. Governatori, P. Noriega, and L. van der Torre. *Normative Multi-Agent Systems*. 04 2013.

[2] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] V. Dignum. Responsible ai and autonomous agents: Governance, ethics, and sustainable innovation. AAMAS '25, page 1–2, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems.

[5] C. D. Fleming, Stephen M. ;Frith. *The cognitive neuroscience of metacognition*. Springer Nature, 2014.

[6] Z. Gou, Z. Shao, Y. Gong, yelong shen, Y. Yang, N. Duan, and W. Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] J. E. L. V. M. G. Ian Kivlichan, Jeffrey Sorensen and P. Culliton. Jigsaw multilingual toxic comment classification, 2020.

[8] D. Kamińska, K. Smółka, and G. Zwoliński. Detection of mental stress through eeg signal in virtual reality environment. *Electronics*, 10(22), 2021.

[9] M. Kim, H. Lee, K. M. Yoo, J. Park, H. Lee, and K. Jung. Critic-guided decoding for controlled text generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4598–4612, Toronto, Canada, 2023.

[10] S. Lee, S. Kim, M. Kim, D. Kang, D. Yang, H. Kim, M. Kang, D. Jung, M. H. Kim, S. Lee, K.-M. Chung, Y. Yu, D. Lee, and J. Yeo. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] S. W. Mercer, M. Maxwell, D. Heaney, and G. C. M. Watt. The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, 21(6):699–705, Dec. 2004.

[13] A. S. Rao and M. P. Georgeff. Bdi agents: From theory to practice. *Proceedings of the First International Conference on Multi-Agent Systems*, 1995.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[15] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

[16] S. Yao, J. Zhao, D. Yu, and et al. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.

[17] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian. Galore: Memory-efficient llm training by gradient low-rank projection, 2024.

2.3 Biased Social Norms of Cooperation in Large Language Models

Biased social norms of cooperation in large language models

Alexandre S. Pires[0000−0002−0648−5702], Laurens Samson, Fernando P. Santos[0000−0002−2310−6444], and Sennay Ghebreab[0009−0007−5788−4635]

Institute of Informatics, University of Amsterdam, Amsterdam, The Netherlands
{a.m.dasilvapires, l.samson, f.p.santos, s.ghebreab}@uva.nl

Abstract. Large language models (LLMs) are increasingly used to evaluate social situations and guide human decision-making. This raises concerns about what rules dictate which behaviors are considered acceptable by LLMs — that is, what their social norms are. Furthermore, it is important to assess what is considered by these social norms, as LLMs might replicate and propagate existing societal biases. Using a prompt dataset portraying interactions between individuals with diverse names, we investigate the social norms implicitly applied by 21 state-of-the-art LLMs when judging cooperative dilemmas within an indirect reciprocity framework, where decisions to cooperate depend on reputations. We find that while LLMs agree on how to judge cooperation and defection against individuals with good reputations, judgments vary substantially when evaluating actions towards ill-reputed individuals. Furthermore, the social norms used by many LLMs are inconsistent across cultural and gender cues stemming from the names of the actors in the interactions. Through evolutionary game-theoretical simulations, we show that these social norms, when used by humans, can limit long-term cooperation. Our results highlight that current LLMs use social norms that are inconsistent within and between families of language models, contain biases, and fail to effectively promote cooperation.

Keywords: Large language models · Cooperation · Social norms · Bias · Indirect reciprocity.

1 Introduction

Large language models (LLMs) are becoming embedded in everyday decision-making contexts, where they are increasingly consulted to interpret social situations, evaluate human actions, and provide guidance to human decision-making [4]. At the same time, these systems raise questions about their broader societal impact [3]. LLM outputs can reflect systematic biases linked to cultural [16], gender [8], or identity cues [6], effectively possessing their own social norms [14]. When these systems are used to assess behavior, these biases may influence collective expectations and contribute to the persistence of social inequalities [17]. Prior studies further indicate that LLMs can affect individuals' ethical judgments

and shared norms [7]. As such, LLMs have their own (biased) social norms while also being capable of influencing human norms. Given the crucial role of social norms in supporting human cooperation by defining what constitutes acceptable behavior [2], it is fundamental to understand if the social norms embedded in LLMs are capable of supporting cooperation, and what biases they may reflect.

Cooperation relies on reputational mechanisms that allow individuals to evaluate partners based on observed behavior. This is known as indirect reciprocity (**IR**), where decisions to cooperate are guided by reputations – opinions regarding the moral standing of others – which are assigned and propagated by observers, shaping subsequent cooperation decisions [9, 18]. In order to effectively assign reputations, humans use shared social norms that determine how actions should be judged given their context [15]. Although norms and reputation dynamics have been extensively studied in human populations [11], the implications of delegating reputation assignments to LLMs remain underexplored [5, 1].

In this work, we investigate three research questions: 1) What social norms do LLMs use in the context of indirect reciprocity? 2) Are these social norms biased depending on cultural or gender identity cues? and 3) Can these social norms sustain human cooperation? To address these questions, we construct a prompt dataset designed to elicit reputational judgments in scenarios illustrating donation games. The dataset comprises descriptions of 43,200 interactions involving agents with varying past reputations, and actor names suggesting various cultures and genders. Each scenario describes a donor's decision to cooperate or defect with a recipient who has a prior reputation (good or bad) in various circumstances (e.g. when asked for money to buy food). We query 21 state-of-the-art LLMs (e.g., GPT-4o, Deepseek R1, Grok 2) to evaluate donors, and translate their responses into probabilistic social norms consistent with standard IR formulations. We then use these social norms in an evolutionary game-theoretical model to study their long-term capacity to sustain cooperation. Our extended study of LLM social norms is available at [12].

2 Methods

We next detail the assessment process of the social norm for a given LLM. First, we outline the construction of the prompt dataset, followed by the method to process responses and aggregate them to define a social norm. The models we test include: GPT-3.5-Turbo and GPT-4o, Qwen 2.5 7B IT and Qwen 2.5 14B IT, Gemma 2 9B IT and Gemma 2 27B IT, Gemini 1.5 Pro and Gemini 2.0 Flash, Mistral Small and Mistral Large, Phi-3.5 Mini IT and Phi-4, Llama 2 7B and Llama 2 13B, Llama 3.1 8B IT and Llama 3.3 70B IT, Claude 3.5 Haiku and Claude 3.7 Sonnet, Grok 2, Deepseek V3 and Deepseek R1 (see [12] for all model details), where "B" is the parameter size of the model in billions, and IT refers to an instruction-tuned version of a model. To ensure reproducibility, all models are queried using a temperature of zero.

Prompt dataset The prompt dataset is constructed by providing LLMs the context of a donation game interaction, and requesting them to output an assessment regarding the reputation of the donor. We make use of second-order social norms, where the reputation of a donor depends on its action (**C** or **D**) and on the reputation of the receiver (**G** or **B**). Each prompt first presents the donor and receiver to the LLM, together with the prior reputation of the receiver, followed by a description of their interaction and the action (help or not help) chosen by the donor. Finally, the LLM is instructed to provide its new reputation, answering exclusively "good" or "bad". To ensure variety in the dataset, we used 5 template prompts, presenting the structure detailed above but varying in phrasing, containing fields (e.g., the donor's action) to be filled. Various possible elements were defined, containing names of different genders and regions to be used for the donor and receiver, the available actions for the donor, as well as different contexts for interactions, such as the donor asking for money or food. All possible combinations of prompts were generated, excluding prompts in which donor and recipient share names, for a total of 43,200 prompts (see [12] for all the details regarding the prompt dataset). Formally, we define $\mathcal{D}'$ to be the dataset composed of all the prompts detailed above, for a given dataset variation. From there, a subset of the dataset is defined as $\mathcal{D}^f = f(\mathcal{D}')$, where f is a filtering function that returns a subset of the dataset that matches some given criteria (e.g., the donor's action). Our final set of datasets, $\mathcal{T}$ is composed of $\mathcal{D}'$ as well as subsets corresponding to different pairs of donor and receiver genders, name regions, and contexts of interaction. Finally, we define $D_{X,Y}$ as the set of prompts in a dataset D where the donor executes action $Y \in \{C, D\}$, and the receiver has a previous reputation $X \in \{G, B\}$.

Norm aggregation After prompting an LLM on the prompt dataset, we next parse each of its answers. For each response, a reputation value $o \in [0, 1]$ is assigned depending on the content of the reply, with 1 and 0 corresponding to assigning the donor a good and bad reputation, respectively. Although our prompts present formatting instructions, not all models adhere to the required format. We parse solely the first paragraph with content of each response, omitting any discussion or justification by the model. If the answer contains "good" and not "bad", 1 is assigned. If it contains exclusively "bad", 0 is assigned. An additional rule is used to account for formatting errors: answers with "neutral" assign 0.5. Any other answer is considered invalid. For each dataset $\mathcal{D} \in \mathcal{T}$, a social norm is defined by evaluating the average value of o, $\bar{o}^{\mathcal{D}}$, of all valid answers at each pair of actions and reputations. Using dataset $\mathcal{D}$, the social norm is given by $d^{\mathcal{D}} = \{\bar{o}^{\mathcal{D}_{G,C}}, \bar{o}^{\mathcal{D}_{G,D}}, \bar{o}^{\mathcal{D}_{B,C}}, \bar{o}^{\mathcal{D}_{B,D}}\}$. In addition, we model the norm variance using a multivariate Gaussian distribution centered at $d^{\mathcal{D}'}$. The covariance matrix for this distribution is estimated as the weighted covariance of $d^{\mathcal{D}}, \mathcal{D} \in \mathcal{T} \setminus \{\mathcal{D}'\}$ weighted by $|\mathcal{D}|$.

Cooperation model We study a finite, well-mixed population of Z adaptive individuals representing humans, following standard formulations of **IR** [13]. In-

teractions take place through repeated donation games in which a donor decides whether to cooperate, **C**, incurring a cost c to provide the recipient with a benefit b ($b > c > 0$), or to defect, **D**, which has neither cost nor benefit. Observers maintain reputational assessments of others, classifying each individual as either **Good (G)** or **Bad (B)**. We consider public reputations, where agents converge to a shared evaluation through gossip. Agent behavior is determined by strategies that condition actions on the recipient's reputation. A strategy is defined as $s = (s_G, s_B)$, where s_G and s_B denote the probability of cooperating with agents perceived as **G** or **B**. At any time, adaptive agents use one of three strategies: *AllC* $(1, 1)$, which always cooperates; *AllD* $(0, 0)$, which always defects; and *Disc* $(1, 0)$, which cooperates only with **G** recipients. Execution errors are included with probability e_e of a cooperation resulting instead in a defection.

Reputational updates follow second-order social norms [13], which evaluate a donor's action in relation to the recipient's reputation. A norm is encoded by a 4-bit vector $d = (d_{G,C}, d_{G,D}, d_{B,C}, d_{B,D})$, giving the probability of assigning a good reputation in each scenario, exactly as the social norms extracted from LLMs. Assessment errors are included as a probability e_a to incorrectly assign a reputation following an observation.

Strategy evolution among adaptive agents follows a birth–death process [15] driven by mutation and social learning. With probability γ, agents adopt a random strategy through mutation. Otherwise, they update their strategy by imitating another individual. This is modeled through pairwise comparison (Fermi rule), whereby imitation likelihood increases with fitness differences between strategies. Population dynamics are represented as a Markov chain in which each state corresponds to a configuration of strategies, and transitions represent changes in the strategy of a single agent. Full derivations are provided in [12]. Cooperation levels are defined as the average proportion of cooperative actions observed in the population over time.

3 Results

Figure 1 summarizes the social norms inferred from each LLM tested. Across models, judgments are largely aligned when donors interact with well-reputed recipients: cooperation is almost universally considered good, whereas refusing to help is considered bad. This pattern is consistent with cooperative norms identified in prior work on **IR** [10]. In contrast, there is little agreement between models when the recipient has a bad reputation. While many models still reward cooperation, they diverge in how they evaluate defection, producing norms distributed along the continuum between Image Score (cooperation is always good, defection is always bad) and Simple Standing (cooperation is good, defecting with bad individuals is also good). A smaller subset of models adopts stricter or more punitive rules, such as Shunning (only cooperation with **G** is good) or Stern Judging (cooperation with **G** or defection with **B** is good), that reward defection against bad individuals over cooperation, whereas others display inconsistent responses. Differences also appear within model families and across

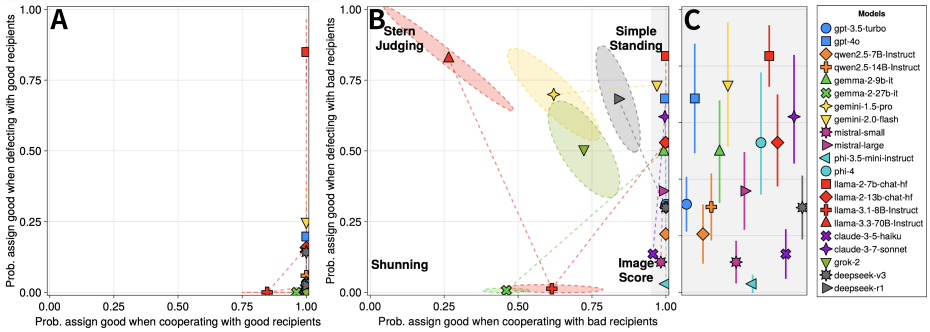


Fig. 1. Social norms inferred from each LLM. Each point represents the mean norm estimated for a given model, where the horizontal axis denotes the probability of assigning a positive reputation after cooperation, and the vertical axis after defection. Ellipses correspond to one standard deviation around the inferred norm, illustrating variability in judgments. Models belonging to the same family share a color and are linked according to version progression and number of parameters. **A**: Evaluations of interactions involving recipients with good reputations. The concentration of points in the lower-right region indicates agreement across models that cooperation with good recipients is good and defection is bad. **B**: Evaluations when recipients have a bad reputation. Many models judge cooperation as good but diverge in how they treat defection, while other models consider cooperation with bad individuals bad. Commonly studied social norms are indicated for reference: Image Score, Shunning, Stern Judging and Simple Standing. **C**: Standard deviation for models that consider cooperation with bad recipients as good (area in **B** shaded in gray).

parameter sizes. Newer and larger models often consider the reputation of the recipient in addition to the action done by the donor. This indicates that model architecture and training regimes can result in vastly distinct social norms.

Because our dataset contains various interaction contexts between actors with different names, the inferred norms exhibit measurable variance, visualized as ellipses in Figure 1. Certain models produce highly stable judgments across prompt variations, whereas others show substantial dispersion, suggesting sensitivity to contextual signals. In Figure 2 we present the resulting level of cooperation across the norm space, highlighting how the social norms of a subset of large language models support distinct levels of cooperation in our evolutionary model. We further decompose the social norm of these LLMs based on the suggested gender (female or male) and region (Western, East Asian or Middle Eastern) of the recipient and donor's names. We observe a notable difference in social norms depending on both these identities, which while small in the space of all possible social norms result in significant differences in cooperation. For example, Grok 2 and Llama 3.1 8B Instruct more often assign lower reputations to female donors cooperating with bad male recipients than the inverse, and these differences are amplified if the female is implied to be of a western region.

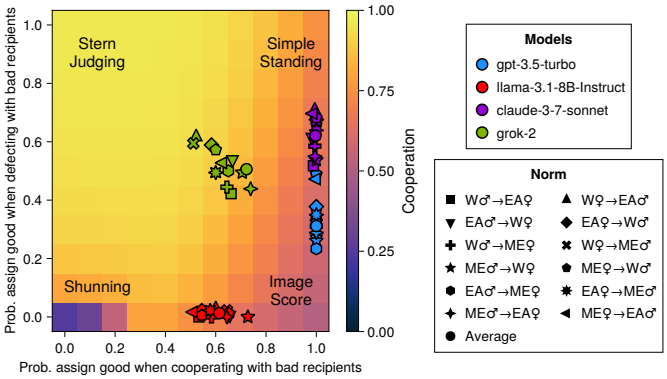


Fig. 2. Prevalence of cooperation across the same space of social norms as Figure 1B. The remaining norm assumes cooperation with good individuals is always good and defection with good individuals is bad. A subset of the norms from different LLMs are overlaid, separated by interactions between donors and recipients of different genders (male, ♂, or female, ♀) and regions (West, W, East Asia, EA, and Middle East, ME). For example, W♂→EA♀ represent the extracted norm used in interactions between a male Western donor interacting with an East Asian female recipient. We observe that the area surrounding Stern Judging achieves the highest level of cooperation, while most models stay between Image Score and Simple Standing. Furthermore, LLMs exhibit substantial gender and cultural biases that result in distinct levels of cooperation between groups. Parameters used: $Z = 100, b/c = 5.0, e_e = e_a = 0.01, \gamma = 0.01, \beta = 1$.

4 Conclusion

We studied the social norms embedded in 21 state-of-the-art LLMs within an indirect reciprocity framework. While LLMs frequently assign good reputations to cooperation with positively reputed individuals, their judgments diverge when evaluating interactions with ill-reputed recipients. These differences span model families, sizes, and versions, with larger and newer models often adopting more context-sensitive norms than smaller or earlier versions. Importantly, these social norms also differ significantly when the names of the actors involved suggest different genders or cultural backgrounds. Using an evolutionary game-theoretical model, we show that these inconsistencies can substantially affect long-term cooperation across society while also propagating biases. These findings highlight the importance of evaluating LLM social norms before deploying them in contexts where they inform human judgments. As such, we urge LLMs to be embedded with concrete, culturally aware, and transparent social norms, while acknowledging the normative challenges this entails. Future work could extend this framework to other potentially more complex social dilemmas, in order to assess to what extent the type of interaction influences LLM judgment. Furthermore, while our work relies on names to convey gender and cultural information, names might also carry other information (social status, religion), which may confound our findings and require new methods to disambiguate.

References

1. Akata, Z., Balliet, D., De Rijke, M., Dignum, F., et al.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (Aug 2020)
2. Alexander, R.: The Biology of Moral Systems. Routledge (2017)
3. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv arXiv:2108.07258 (2021)
4. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–45 (2024)
5. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative ai: machines must learn to find common ground. Nature **593**(7857), 33–36 (2021)
6. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. Computational Linguistics **50**(3), 1097–1179 (2024)
7. Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M.: Co-writing with opinionated language models affects users' views. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems pp. 1–15 (2023)
8. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. Proceedings of the ACM Collective Intelligence Conference **CI '23**, 12–24 (2023). https://doi.org/10.1145/3582269.3615599, https://doi.org/10.1145/3582269.3615599
9. Nowak, M.A., Sigmund, K.: Evolution of indirect reciprocity. Nature **437**(7063), 1291–1298 (2005)
10. Ohtsuki, H., Iwasa, Y.: The leading eight: social norms that can maintain cooperation by indirect reciprocity. Journal of Theoretical Biology **239**(4), 435–444 (2006)
11. Okada, I.: A Review of Theoretical Studies on Indirect Reciprocity. Games **11**(3), 27 (Jul 2020)
12. Pires, A.S., Samson, L., Ghebreab, S., Santos, F.P.: How large language models judge and influence human cooperation. arXiv preprint arXiv:2507.00088 (2025)
13. Santos, F.P., Santos, F.C., Pacheco, J.M.: Social norms of cooperation in small-scale societies. PLoS Computational Biology **12**, e1004709 (2016)
14. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A., Kersting, K.: Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence **4**(3), 258–268 (2022)
15. Sigmund, K.: The Calculus of Selfishness. Princeton University Press (2010)
16. Tao, Y., Viberg, O., Baker, R.S., Kizilcec, R.F.: Cultural bias and cultural alignment of large language models. PNAS Nexus **3**(9), pgae346 (2024)
17. Wang, A., Morgenstern, J., Dickerson, J.P.: Large language models that replace human participants can harmfully misportray and flatten identity groups. Nature Machine Intelligence pp. 1–12 (2025)
18. Wu, J., Balliet, D., Van Lange, P.A.: When does gossip promote generosity? indirect reciprocity under the shadow of the future. Social Psychological and Personality Science **6**(8), 923–930 (2015)

3 Fairness, Institutions, and Strategic Public Decision-Making

3.1 Long-term Effects of Fairness Metrics on Population Dynamics

Long-term Effects of Fairness Metrics on Population Dynamics

Mirthe Dankloff[1][0009−0009−4439−5749], Yining Yuan[2][0009−0000−3093−5526], Nirav Ajmeri[2][0000−0003−3627−097X], and Vahid Yazdanpanah[3][0000−0002−4468−6193]

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands m.e.dankloff@vu.nl
University of Bristol, Bristol, United Kingdom
{yining.yuan,nirav.ajmeri}@bristol.ac.uk
University of Southampton, Southampton, United Kingdom
V.Yazdanpanah@soton.ac.uk

Abstract. Algorithmic fairness is often treated as a static property, overlooking that individuals may disengage from systems they perceive as unfair. We introduce a dynamic notion of perceived fairness in a lending scenario that captures how repeated unjust denials and observation of peer outcomes can drive applicants to opt out. Through a multi-agent simulation framework on synthetic data, we measure how different fairness metrics affect long-term population retention and feature dynamics. Our results show that without fairness constraints, apparent fairness improvements arise from the selective opt-out of disadvantaged applicants (survivorship bias). Demographic parity and equal opportunity reduce immediate retention disparities but do not guarantee long-term fairness; demographic parity, in particular, overcorrects participation dynamics, accumulating long-term unfairness. We compare this against a causal fairness model that achieves a balanced retention rate and lower long-term unfairness. Our findings highlight the need to assess long-term fairness in settings with endogenous participation, where individual decisions are shaped by perceived fairness and peer effects, beyond static fairness constraints.

Keywords: Long-Term Fairness · Social Influence · Agent-Based Simulation.

1 Introduction

Algorithmic systems are increasingly used for resource allocation decisions such as lending and other public services, making fair treatment between demographic groups a critical concern [2,6,9]. Fairness metrics (e.g., demographic parity) have been proposed to equalize the outcomes between groups through statistical constraints [1,6]. However, prior research has shown that applying such metrics over time can produce unintended effects, such as overlending to disadvantaged groups or amplifying disparities, highlighting the need to study long-term fairness dynamics [3,8,12]. Existing research on long-term fairness primarily model feature dynamics, examining how decisions affect future applicant qualifications

(e.g., credit scores) [3,8,12]. These studies adopt a decision-maker's perspective, assuming static participation, and overlooking that applicants may disengage from systems they perceive as unfair. Spillover effects, where decisions about some applicants influence the behavior of others, are also often ignored [10,11]. Satisfying group-level fairness is thus no guarantee that the outcomes are perceived as fair by individual applicants [4,15].

Against this background, we ask: *to what extent do lending decisions lead to fairer outcomes over time when considering perceived fairness, feature dynamics, and population dynamics?* We address this question by evaluating long-term fairness under endogenous participation driven by perceived fairness and peer effects.

Building on work on *perceived fairness* in psychology and algorithmic decision-making [13,14], we introduce a variation of this notion that captures both individual experience and social observation within a multi-agent simulation. *We define perceived (un)fairness as the extent to which applicants perceive outcomes as favorable to themselves and their peer group.* We introduce a fixed social network through which applicants can observe peer outcomes. Applicants may opt out following repeated unjust rejections or when they observe that their demographic group is denied loans more often than other group members. In the lending context, the applicant's opt-out behavior may refer to switching to competing lenders, which could reduce the amount of customers and representative data for the bank's future decision models. This creates population dynamics with opt-out behavior resulting from perceived unfairness. To evaluate these dynamics, we propose retention rate and retention disparity among demographic groups as metrics that capture differential opt-out.

We evaluate our framework in a synthetic lending setting [7], comparing demographic parity, equal opportunity, and causal fairness. Our results show that perceived unfairness can induce selective opt-out among disadvantaged groups, reducing measured unfairness through population change rather than improved treatment. This "survivorship bias"[5] also occurs in baseline conditions without fairness metrics. While demographic parity and equal opportunity reduce immediate retention disparities, they do not guarantee long-term fairness. In contrast, the causal fairness model [7] achieves balanced retention and lower long-term unfairness. These findings highlight trade-offs between accuracy, retention, and long-term fairness that are not observable in static settings. Our work emphasizes the need to jointly evaluate fairness constraints with endogenous participation dynamics to ensure that fairness interventions serve the citizens they are designed to protect.

2 Method

In this section, we describe the lending environment and social network (2.1), the perceived fairness state and opt-out behavior (2.2), the bank's decision model (2.3), and the evaluation metrics (fairness and retention) (2.4).

2.1 Bank Loan Simulation and Social Network

Each applicant i has a binary protected attribute $S \in \{0, 1\}$ (e.g., age), a vector of dynamic features $\mathbf{X}_t \in \mathbb{R}^d$ (e.g., credit risk score at time t), and a ground-truth label $Y_{i,t} \in \{-1, 1\}$ indicating whether the applicant can pay back ($Y_{i,t} = 1$) or default ($Y_{i,t} = -1$). At each time step, the lender makes a binary loan decision $D_{i,t} \in \{-1, 1\}$, where $D_{i,t} = 1$ denotes approval and $D_{i,t} = -1$ denial. Each applicant then chooses the action $A_{i,t} \in \{0, 1\}$ to apply ($A_{i,t} = 1$) or opt out ($A_{i,t} = 0$) in the lending process. Our synthetic datapoint has $|\mathbf{X}_1| = 2$ non-protected features sampled from S-specific Gaussian distributions and Y sampled from a ground-truth model. The detailed generation process for further steps is given in Appendix A.1.

Instead of simulating the behavior of individuals in isolation, we model how fairness is mediated through sequential social interactions. Applicants are embedded in a fixed, undirected social network. This can take spillover effects into account, where one applicant's rejection can influence another applicant's decision to participate [11]. Each applicant is connected to a predefined set of neighbors $\mathcal{N}_i$ with degree $n = 10$, where 80% of neighbors share the same protected attribute ($S_i = S_j = s, j \in \mathcal{N}_i$) and 20% of neighbors belong to the other demographic group ($S_i = s, S_j = s', j \in \mathcal{N}_i$). We choose this ratio to reflect the tendency for individuals to connect more often with similar individuals.

2.2 Perceived Fairness and Opt-out Behavior

Research on perceived fairness highlights two dimensions: whether individuals view their own outcomes as just and whether they perceive the process as legitimate relative to others' outcomes [13,14]. Inspired by this, we model perceived fairness as a deterministic belief state formed by applicants, unobservable by the lender, and separate from the fairness metrics used for evaluation. Furthermore, we assume that applicants have full knowledge of their own application details and history but observe only the decision outcomes of their peers.

At each time step t, an applicant i observes the outcomes of its connected neighbors $\mathcal{N}_i$ to form a peer-outcome signal $O_{i,t} \in \{0, 1\}$. $\mathcal{N}_{i,t}^{\text{act}} \subseteq \mathcal{N}_i$ denotes the subset of neighbors who are active at time t. The applicant computes the observed rejection rate for each group as $r_s = \frac{\sum_{j \in \mathcal{N}_{i,t}^{\text{act}},\, S_j = s} \mathbb{1}[D_{j,t} = -1]}{|\{j \in \mathcal{N}_{i,t}^{\text{act}} : S_j = s\}| + \epsilon}$. In this equation, ϵ in the denominator is a small constant introduced to prevent division by zero when agent i has no active neighbors of group s. The peer-outcome signal $O_{i,t} = 1$ if $r_{S_i} > r_{1-S_i}$, otherwise $O_{i,t} = 0$. This signal reflects whether the applicant's in-group experiences a higher rejection rate than the out-group, generating a social spillover effect that influences perceived fairness. Let $P_{i,t} = D_{i,t-1}$ store the last decision received from the bank.

We define perceived unfairness as $U_{i,t} = \mathbf{1}[D_{i,t} = -1 \wedge ((P_{i,t} = -1 \wedge Y_{i,t} = 1) \vee O_{i,t} = 1)]$. The first condition represents an applicant who can pay back but is unjustly denied a loan in two consecutive time steps, while the second captures if the in-group rejection rate exceeds the out-group rejection rate for a

rejected agent. The agent opts out $A_{i,t+1} = 0$ if it perceives unfairness ($U_{i,t} = 1$), or because it is correctly identified as fraudulent ($D_{i,t} = Y_{i,t} = -1$). Figure 1 demonstrates an example of three consecutive simulation steps. A_5 opts out at t_2 because it was correctly identified as fraud, which does not trigger perceived unfairness. A_3 opts out at t_3 because it was denied a loan in two consecutive time steps, and A_1 will opt out at the next step (t_4) due to spillover effects.

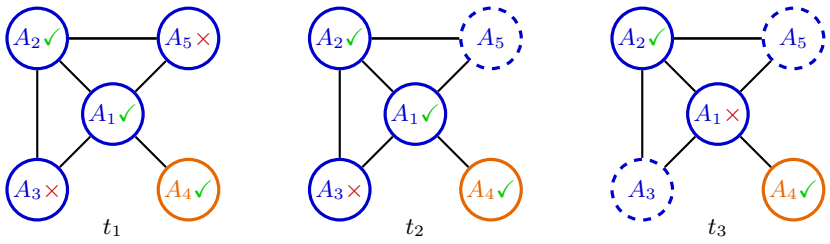


Fig. 1: Blue and orange agents denote different protected groups. Dashed circles represent $A_{i,t} = 0$. ✓ means application approved, × means application rejected.

2.3 Bank's Decision Model

We evaluate three fairness constraints. *Demographic Parity (DP)* requires equal approval rates between groups defined by the protected attribute, *Equal Opportunity (EO)* requires equal true positive rates between groups, and *Long-term Counterfactual Fairness (LCF)* requires that an agent's approval probability at a future horizon t^* be invariant to the protected attribute. The bank's decision model is optimized using Algorithm 1. The bank optimizes prediction accuracy, while setting either DP, EO or LCF as upper-bounded constraints. These are compared against a baseline model without fairness constraints.

2.4 Evaluation metrics

We evaluate fairness from two perspectives. From the lender's perspective, we measure *accuracy* (true positive rate at each step) and *t-step (long-term)*. We adopt the t-step definition from [7], which captures whether cumulative feature trajectories create divergent outcomes between demographic groups over time, where 0 is perfectly fair. To capture the population dynamics, we introduce three retention metrics: the *retention rate* $Rt = \frac{1}{N}\sum_{i=1}^{N} A_{i,t}$, the *retention disparity* $\Delta t = |R_t^{S=0} - R_t^{S=1}|$ which captures the differential opt-out between the groups, and the *relative retention ratio* $\rho = \frac{R_t^{S=0}}{R_t^{S=1}}$ where values below 1 indicate disproportionate opt-out for the protected group.

Algorithm 1 Bank's Decision Model

Input: $X = [S, \mathbf{X}] \in \mathbb{R}^{n \times (d+1)}, Y \in \{-1, 1\}^n$
Parameters: Weight vector $\mathbf{w} = (w_s, w_1, \ldots, w_d)$, intercept c, logistic function σ
for each applicant i at time t **do**
$p_{i,t} \leftarrow \sigma(w_s s_i + \sum_{j=1}^{d} w_j X_{i,t}^j + c)$,
$D_{i,t} \leftarrow 2\,\mathbf{1}[p_{i,t} > 0.5] - 1$
end for
Solve: $\min_{\mathbf{w},\, c}\ \mathcal{L}(\mathbf{w}, X)$ subject to $\mathcal{C}(\mathbf{w}, X) \leq \tau$
Optional Fairness constraints:
DP: $| P(D_t{=}1 \mid S{=}0) - P(D_t{=}1 \mid S{=}1) | \leq \tau$
EO: $| P(D_t{=}1 \mid Y_t{=}1, S{=}0) - P(D_t{=}1 \mid Y_t{=}1, S{=}1) | \leq \tau$
LCF: $| P(D_{t^*} = 1 \mid S = s, X = x) - P(D_{t^*} = 1 | S = s', X = x) | \leq \tau$
return Optimized $\mathbf{w}^*, c^*$

3 Preliminary Results

Accuracy The baseline achieves the highest accuracy ($t_5 = 0.734$) across all time steps compared to the fairness-aware approaches. EO has a slightly lower accuracy ($t_5 = 0.690$) followed by LCF ($t_5 = 0.683$) and DP ($t_5 = 0.685$), suggesting a modest accuracy trade-off when optimizing with fairness constraints. Detailed results for comparisons between different methods that evaluate accuracy, long-term unfairness, retention rate, and retention disparity are provided in the Appendix A.2.

Long-term fairness t-step Figure 2 shows that LCF achieves the lowest cumulative unfairness across all time steps, demonstrating that incorporating multi-step counterfactual reasoning into the fairness constraint [7] remains effective even under population dynamics. In contrast, DP and EO constraints enforce group fairness at each step but do not account for the cumulative feature trajectories induced by their decisions. The resulting mismatch accumulates over time, resulting in both DP and EO exceeding the long-term unfairness of the baseline model without population dynamics, consistent with findings in [8]. When comparing the simulation with and without population dynamics in Figure 2, the baseline and EO exhibit higher unfairness in a static setting when the full population is retained. DP shows a higher unfairness with population dynamics from step 4. LCF shows almost no difference between the settings, demonstrating that long-term fairness does not depend on selective opt-out of disadvantaged groups.

Retention rates Figure 3 shows that the baseline condition exhibits a severe retention disparity between groups, with the protected group ($S = 0$) leaving at a higher rate. This indicates a survivorship bias where long-term fairness seems to improve because disadvantaged applicants left the system [5]. The relative retention ratio (right axis) confirms this: the baseline ratio decreases and remains below 1.0. While DP and EO reduce this gap, they do so inconsistently: EO narrows but does not eliminate the imbalance, while DP reverses the disparity so that the unprotected group leaves faster (relative retention ratio above 1.0).

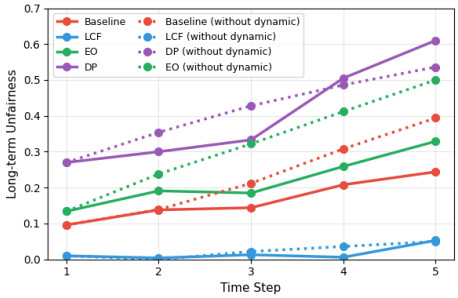


Fig. 2: Long-term unfairness with (solid) and without (dotted) population dynamics for each fairness metric. LCF maintains near-zero unfairness in both settings; DP has the highest long-term unfairness.

The LCF model achieves near equal retention with a ratio close to 1.0 throughout while maintaining counterfactual fairness.

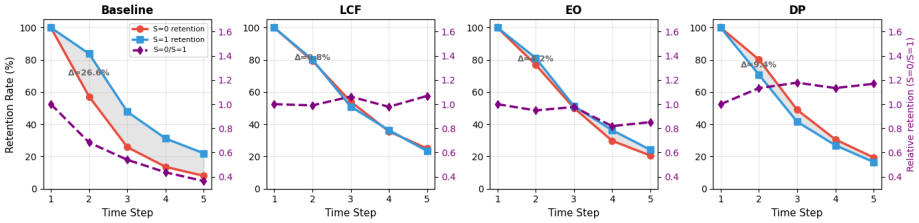


Fig. 3: Retention rates per group (left-axis) for $S = 0$ (red) and $S = 1$ (blue); relative retention ratio (right axis) for $S = 0/S = 1$ (purple dashed), values below 1.0 indicate disproportionate opt-out for $S = 0$. Baseline shows severe retention disparity ($\Delta = 26.6\%$), LCF achieves near parity ($\Delta = 0.8\%$) and DP reverses disparity.

4 Conclusions and Future Work

In this work, we evaluate long-term fairness under endogenous participation driven by perceived fairness and peer effects. We introduce perceived fairness as an internal belief state that captures past experience and social observation. Unlike prior studies that model feature evolution but assume static populations [3,8,12], our framework integrates population dynamics driven by perceived fairness, exposing trade-offs between accuracy, short-term fairness, and long-term fairness. While the unconstrained baseline achieves the highest accuracy, it shows high retention disparities, driven by selective opt-out of the protected group.

Demographic parity and equal opportunity reduce the immediate retention disparity, but do not guarantee long-term fairness. Demographic parity overcorrects retention in earlier time steps, resulting in reverse long-term disparities. In contrast, the causal fairness model consistently achieves low long-term unfairness with balanced retention rates between groups. Our findings emphasize the importance of a citizen-centric perspective in evaluating algorithmic decision-making: when fairness interventions fail to account for how citizens perceive and react to decisions, this may inadvertently drive the most vulnerable individuals out of essential public services, undermining the fairness goals they aim to achieve.

To build on this workshop paper, we plan to validate our framework on real-world datasets and study additional fairness metrics (e.g., predictive equality). The current results are derived under the assumption of a fixed network degree and group-mixing ratios. Future work will address these constraints by performing a sensitivity analysis on neighborhood size and the proportion of connections between demographic groups. Moreover, we can extend this work by generalizing the two-step denial memory to variable lengths, modeling dynamic social networks where peer groups evolve over time, and incorporating positive spillover effects arising from favorable outcomes. Additional directions include introducing competing lenders to capture push-pull dynamics and modeling the lender as a learning agent that optimizes long-term objectives, such as retention and perceived fairness.

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning. Recommender systems handbook **1**, 453–459 (2020)
2. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
3. D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., Halpern, Y.: Fairness is not static: deeper understanding of long term fairness via simulation studies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 525–534 (2020)
4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)
5. Gupta, P., MacAvaney, S.: On survivorship bias in ms marco. In: proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval. pp. 2214–2219 (2022)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
7. Hu, Y., Zhang, L.: Achieving long-term fairness in sequential decision making. Proceedings of the AAAI Conference on Artificial Intelligence **36**(9), 95499557 (Jun 2022). https://doi.org/10.1609/aaai.v36i9.21188
8. Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M.: Delayed impact of fair machine learning. In: International Conference on Machine Learning. pp. 3150–3158. PMLR (2018)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
10. Mehrotra, A., Sachs, J., Celis, L.E.: Revisiting group fairness metrics: The effect of networks. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–29 (2022)
11. Narayanan, A.: What if algorithmic fairness is a category error? Contemporary Debates in the Ethics of Artificial Intelligence pp. 77–95 (2025)
12. Rateike, M., Valera, I., Forré, P.: Designing long-term group fair policies in dynamical systems. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 20–50 (2024)
13. Tyler, T.R.: Social justice: Outcome and procedure, 35 intl j. Psychol **117**, 119–20 (2000)
14. Wang, R., Harper, F.M., Zhu, H.: Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 114. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376813, https://doi.org/10.1145/3313831.3376813
15. Zhou, W., et al.: Group vs. individual algorithmic fairness (2022)

A Appendix

A.1 Synthetic Dataset

We generate a 5-step synthetic dataset with 4000 agents for training and 1000 for testing. At $t = 1$, S is sampled with equal probability, and $\mathbf{X}_1$ is sampled from S-specific Gaussian distributions. The true repayment is sampled from a ground-truth decision model $Y_t = \sigma(h_{\theta_*}(\cdot))$, where $\sigma(\cdot)$ is the sigmoid function, $h_{\theta_*}(\cdot)$ is a fixed mapping of the probability of whether an applicant would default $P(Y_t) = \sigma(h_{\theta^*}(\mathbf{X}_t, S))$, $Y_t \sim 2 \cdot \text{Bernoulli}(P(Y_t)) - 1$. D_t is sampled from a separate $\sigma(h_{\theta_*}(\cdot))$ as $D_t \sim 2 \cdot \text{Bernoulli}(P(D_t)) - 1$. $\mathbf{X}_{i,t+1}$ is generated according to the update rule below:

$$\mathbf{X}_{i,t+1} = \begin{cases} \mathbf{X}_{i,t} - \lambda \cdot \theta_t + b & D_{i,t} = 1, Y_{i,t} = -1 \\ \mathbf{X}_{i,t} + \lambda \cdot \theta_t + b & D_{i,t} = 1, Y_{i,t} = 1 \\ \mathbf{X}_{i,t} + b & D_{i,t} = -1 \end{cases} \tag{1}$$

where λ controls the sensitivity of the update to the predicted decisions, and $b = S \cdot b_1 + (1 - S) \cdot b_0$ is a small increment at each time step. The parameters are set as $\lambda = 0.5$, $b_0 = 0.2$, $b_1 = 1.0$.

A.2 Detailed Results Table

Table 1 lists the results of the evaluation metrics in our simulation.

Table 1: Accuracy, long-term unfairness (t-step), retention rate, and retention disparity across fairness metrics over 5 time steps.

Metric	Fairness Metric	t_1	t_2	t_3	t_4	t_5
Accuracy (%)	Baseline	70.4	73.5	72.5	71.8	73.4
	LCF	68.5	70.1	69.9	68.7	68.3
	EO	68.6	69.2	69.9	68.6	69.0
	DP	67.7	67.4	68.1	68.4	68.5
t-step	Baseline	0.096	0.138	0.144	0.208	0.244
	LCF	0.010	0.004	0.013	0.006	0.053
	EO	0.134	0.191	0.185	0.259	0.329
	DP	0.270	0.300	0.333	0.505	0.610
Retention (%)	Baseline	100	70.5	36.8	22.4	15.0
	LCF	100	79.8	52.6	36.0	24.2
	EO	100	79.1	50.8	33.1	22.4
	DP	100	75.5	45.3	28.8	18.0
Disparity (%)	Baseline	0.0	26.6	22.0	17.6	14.0
	LCF	0.0	0.8	3.2	0.8	1.6
	EO	0.0	4.2	1.2	6.6	3.6
	DP	0.0	9.4	7.4	3.6	2.8

3.2 Strategic Exploitation of Placement Guarantees in Random Serial Dictatorship: Modelling the Amsterdam School Choice System

Strategic Exploitation of Placement Guarantees in Random Serial Dictatorship: Modelling the Amsterdam School Choice System

Mayesha Tasnim*[1], Joseph Trevorrow*[2], and Nirav Ajmeri[2]

[1] University of Amsterdam, Amsterdam, Netherlands m.tasnim@uva.nl
[2] University of Bristol, Bristol, BS8 1UB, United Kingdom
{j.trevorrow,nirav.ajmeri}@bristol.ac.uk

Abstract. School choice mechanisms, such as the well-known Deferred Acceptance algorithm, are often modified to meet practical constraints. One such modification is a *placement guarantee*, which stipulates that students who rank at least l schools and are not placed in the main round are given access to one of the schools on their submitted list by increasing school capacities (for example by 4% in Amsterdam). While intended to provide safety to students, this rule changes incentives and can lead to strategic behaviour. We study an Amsterdam-inspired school choice model that combines Random Serial Dictatorship (RSD) with a placement guarantee. We make three contributions. First, we formalise a simple but behaviourally relevant strategic response to placement guarantees: students with sincere preferences over at most $k < l$ schools report their top k schools truthfully and then fill the remaining $l - k$ positions with highly demanded schools that are common knowledge to be heavily oversubscribed. Second, we show that, under mild assumptions, this strategy weakly dominates truthful reporting. Third, we use an agent-based simulation framework to show how adoption of this strategy can collectively make the placement guarantee infeasible when the number of guarantee-eligible students exceeds the additional capacity created by the policy. This work is meant as a decision support for policymakers who want to predict the real-world impact of such modified school choice systems.

Keywords: Agent-Based Modelling · School Choice · Incentives.

1 Introduction

Centralised school choice systems are widely used to assign students to secondary schools in a transparent way. A common benchmark is the Deferred Acceptance (DA) algorithm, which produces stable allocations and incentivizes students to report preferences truthfully [4, 12, 9]. Random Serial Dictatorship (RSD) is another widely used mechanism when schools have no preferences over students;

* These authors contributed equally to this work.

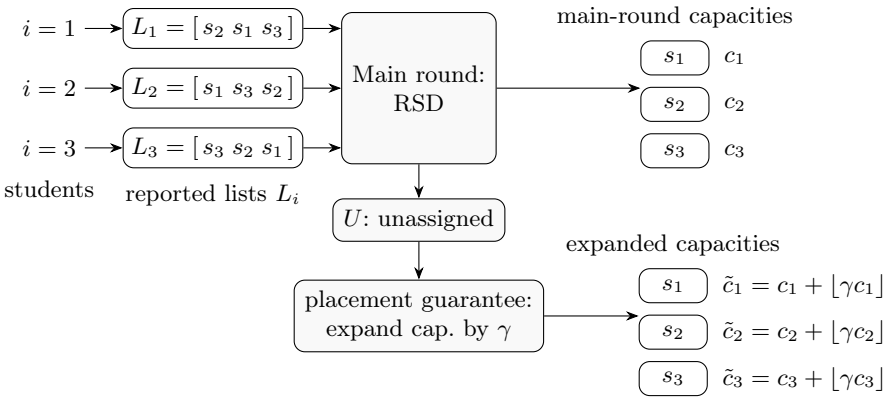


Fig. 1: Conceptual model of school placement under RSD with a placement guarantee. Students submit reported lists L_i, are first assigned via RSD with capacities c_s, and unassigned students enter the placement guarantee stage with expanded capacities $\tilde{c}_s$.

under suitable conditions it also incentivizes truthfulness and is ex-post efficient for the reported preferences [3]. In practice, however, many cities adopt variants with random tie-breaking, priority categories, and ad-hoc policy fixes motivated by local constraints and political objectives [8, 6], which can substantially alter both efficiency and incentives [2, 5].

In Amsterdam, secondary school places are allocated using a procedure that can be viewed as RSD combined with a *placement guarantee*. Students submit a ranked list of at least l schools and are first assigned via RSD with fixed capacities and a random order over students. If a student who listed at least l schools remains unassigned after this main round, the system promises a place at one of the schools on their list, implemented by increasing capacities at those schools by a small fraction (currently around 4%) [13]. Communication from policymakers emphasises that it is safe to list schools truthfully and that the guarantee protects students from ending up without a place. Parents, however, adapt their behaviour to the presence of the placement guarantee [11, 14]. Many families have sincere preferences over fewer than l schools; [3] the remaining positions must be filled with schools they know less well. Filling out the list with merely acceptable schools increases the risk of being placed at an undesired school in the main round, whereas filling it with extremely popular schools that are effectively unreachable raises the chance of remaining unassigned and thus becoming eligible for the guarantee.

This paper asks the following questions. First, when an RSD-based school choice mechanism is supplemented with a placement guarantee, can the guarantee itself create incentives for strategic reporting? Second, under what conditions does this strategic reporting make the guarantee infeasible? We formalise this

[3] Based on a report on school preferences in Amsterdam [10]

situation in a simple model that captures key features of the Amsterdam system. We study a strategy in which students who have sincere preferences over at most $k < l$ schools report these k schools truthfully at the top of their list and then fill the remaining $l-k$ positions with highly demanded schools that are known to be heavily oversubscribed. Intuitively, this either leads to assignment to one of the top k schools in the main round or, failing that, leaves the student unassigned and therefore eligible for the placement guarantee, which creates extra capacity at their preferred schools. Under mild assumptions, we show that this strategy weakly dominates truthful reporting for such students. We also develop a simulation framework in which agents follow either truthful or strategic reporting rules under RSD with a placement guarantee. Using synthetic preference profiles, we vary two key parameters: the number k of sincere preferences per student and the fraction α of strategic agents. We find combinations of parameters where the number of unassigned guarantee-eligible students systematically exceeds the additional seats created, making it impossible to honour the placement guarantee for all students; in these cases, some students cannot be allocated to any school on their list.

2 Model

We consider a finite set N of n students and a set S of m schools. Each school $s \in S$ has base capacity $c_s \in \mathbb{N}$, and each student $i \in N$ has a strict true preference order $\succ_i$ over S. Students submit a reported list L_i, an ordered list of distinct schools in S. There is a minimum list length $l \geq 1$: a student is eligible for the placement guarantee if and only if $|L_i| \geq l$.

In the main round, allocation is performed by Random Serial Dictatorship (RSD) on reported preferences. A permutation π of N is drawn uniformly at random. Starting from capacities c_s, students are considered in the order of π; when student i is called, they are assigned to the highest-ranked school in L_i with remaining capacity. If no school in L_i has remaining capacity, student i remains unassigned. Let $a_i \in S \cup \{\varnothing\}$ denote the main-round outcome for i, and define $U = \{i \in N : a_i = \varnothing \text{ and } |L_i| \geq l\}$ as the set of unassigned students who qualify for the placement guarantee.

The placement guarantee is modelled as a uniform capacity expansion by a factor $\gamma \geq 0$ on all schools: $\tilde{c}_s = c_s + \lfloor \gamma c_s \rfloor, \qquad s \in S$, so the total additional capacity is $C^{\text{extra}} = \sum_{s \in S} \lfloor \gamma c_s \rfloor$. In the guarantee stage, seats allocated in the main round are fixed and only students in U are reconsidered. If $|U| > C^{\text{extra}}$, the guarantee cannot be honoured for all eligible students. A step-by-step description of the resulting allocation procedure is provided in Appendix A.

We distinguish truthful and strategic preference reporting. A truthful student reports a list L_i that coincides with the top $|L_i|$ positions of $\succ_i$. We assume that each student has sincere preferences over at most k schools, with $1 \leq k \leq l$, and is indifferent among all other schools. A fraction $\alpha \in [0, 1]$ of students are strategic, the remaining fraction $1 - \alpha$ are truthful, and we use (k, l, α, γ) as the main parameters of interest. For the analytical results, we assume that there

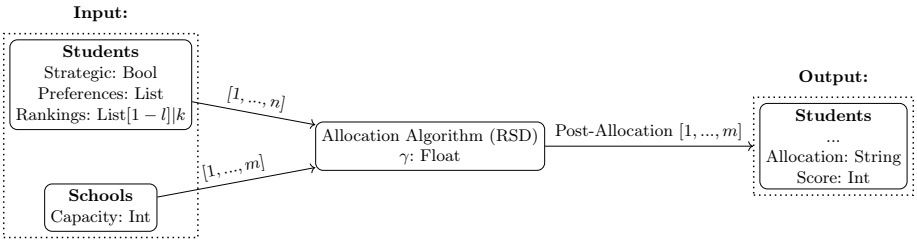


Fig. 2: Pipeline of the simulation environment.

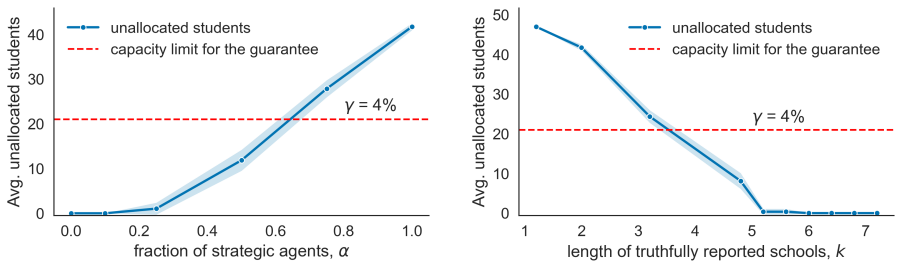
exists a set $P \subseteq S$ of popular schools with $|P| \geq l - k$ that attract a large share of demand and have limited capacity, and that it is common knowledge which schools belong to P. We focus on a simple strategic behaviour in which students list their top k schools truthfully and use popular schools from P to fill the remaining positions up to l. Appendix B formalises this strategy and provides the sketch of a best response.

3 Simulation Environment

We study the mechanism in Section 2 using an agent-based model (ABM) implemented in Python on the PettingZoo library [15]. A pipeline diagram of the environment is shown in Figure 2. The environment contains n students and m schools with fixed base capacities. In each time-step, every student submits a ranked list of length l, the mechanism runs RSD with a placement guarantee as defined in Section 2, and each student receives an allocation and a rank-based score. Agents follow one of two reporting strategies: (i) *truthful* agents report the top l schools of their true preference order; (ii) *strategic* agents use the top-k truthful strategy from Section 2, reporting their sincere top k schools followed by $l-k$ popular schools, where the set and ordering of popular schools are assumed common knowledge.

For a given experiment, we fix all environment parameters and run the same allocation problem 10,000 times to average over the randomness in the RSD order; school capacities, student preferences, and the popular set remain fixed, and we report averages over time-steps. The ABM assumes: (i) stationary agents, with true preferences fixed across time-steps and no learning or adaptation; (ii) baseline sincerity, i.e., truthful agents report their true preferences and strategic behaviour arises only through the specified reporting rules; (iii) random order, with the RSD order drawn from Python's pseudo-random number generator and fixed seeds across experiments for comparability; and (iv) exogenously specified popular schools, known to all agents.

We run experiments using synthetic preferences modelled after anonymised school preference data from the Amsterdam School Boards Association (OSVO) [1]. In this synthetic setting, we consider $n = 100$ students, $m = 20$ schools, and list length $l = 8$, and generate preferences using a popularity-skewed procedure in which a small subset of schools receives higher weight and thus a larger share



(a) With varying α (b) With varying k

Fig. 3: Average number of unassigned students across simulation runs, varying the fraction of strategic agents, α, and number of sincerely preferred schools, k, respectively. The dashed red line marks the total extra capacity from the placement guarantee.

of rank-1 demand. Our main metrics are: (i) the number of guarantee-eligible students, defined as unassigned students after the main RSD round, and (ii) the *guarantee overrun*, defined as the probability that this number exceeds the total additional capacity C^{extra} created by the placement guarantee.

4 Experimental Results

We study how the mechanism behaves as we vary the fraction of strategic agents α and the number of sincere preferences k. Unless stated otherwise, we use the synthetic environment with $n = 100$, list length $l = 8$, and a fixed set of popular schools. For each parameter setting we compute (i) the average number of unassigned students after the main round and (ii) the guarantee overrun probability $\Pr(|U| > C^{\text{extra}})$.

First, we fix k, l, and $\gamma = 0.04$ (as in Amsterdam) and vary the fraction $\alpha \in [0, 1]$ of strategic agents. Figure 3a shows that the average number of unassigned students increases as the fraction of strategic agents α grows. The dashed red line indicates the total additional capacity created by the placement guarantee. When more than 60% of students apply the strategy, the number of guarantee-eligible students exceeds the available capacity extension, and the guarantee cannot be honoured for all eligible students.

Second, we fix $\alpha = 1$ (all agents strategic), keep l and γ fixed, and vary $k \in \{1, \ldots, l-1\}$ as $k = 0$ is substantively unrealistic and $k = l$ corresponds to lists with no strategic padding. Lower values of k mean that students have fewer sincere schools and therefore fill more of their list with popular padding schools. Conversely, when k is close to l, preferences are mostly sincere. Figure 3b shows the values of k for which there are more unassigned students than the guarantee's capacity extension allows. When $k \leq 3$, a large share of each list consists of effectively unreachable popular schools, and the guarantee is repeatedly pushed beyond its available capacity.

5 Discussion and Future Work

We analysed an allocation procedure that combines RSD with a placement guarantee, inspired by the Amsterdam secondary school system. We formalised a simple reporting strategy in which students list their top k schools truthfully and then pad the remainder of the list with popular schools, and showed that under mild assumptions this strategy weakly dominates truthful reporting over the full list. Using an agent-based simulation, we found that when such behaviour becomes widespread, the number of guarantee-eligible students often exceeds the available extra capacity, so the guarantee cannot be honoured for all students who satisfy the list-length requirement.

The present experiments use a deliberately small synthetic environment with $n = 100$, $m = 20$, and $l = 8$ in order to isolate the incentive mechanism rather than provide a calibrated prediction for Amsterdam. Scaling to Amsterdam-sized cohorts is conceptually straightforward, since each run consists of sorting students by a random order and checking capacities along submitted lists; the main challenge is calibration, including realistic capacities, preference distributions, perceived popular schools, and levels of strategic behaviour. The qualitative pattern may also be sensitive to l, γ, and the definition of the popular set, which should be studied systematically in future work.

The placement guarantee is communicated as a promise that students who list at least l schools will receive a place at one of them, but our results show that under plausible levels of strategic reporting this promise cannot always be met. When many families adopt the top-k truthful strategy, the guarantee becomes infeasible, and some students remain unassigned despite complying with the rule. This gap between the stated guarantee and the mechanism's actual capabilities risks undermining trust in both the allocation procedure and the responsible institutions. From a design perspective, the combined mechanism no longer preserves the straightforward truthfulness incentives associated with standard RSD, because the fallback stage creates an additional outcome that families may rationally target. Robust design therefore requires stress-testing placement guarantees under plausible strategic responses before public communication frames truthful listing as safe.

Our analysis focuses on incentives and feasibility rather than distributional impacts. Díaz et al. [7], for instance, simulate school choice under information asymmetries using Chilean enrolment data and show that simplified performance signals help low-income families but do not remove inequalities due to information and geography. Future work includes a more complete analysis of best responses and equilibrium behaviour under RSD with placement guarantees, experiments calibrated to real preference and allocation data from Amsterdam, and the design and evaluation of alternative guarantee rules.

Acknowledgments. JT is supported by the UK Research and Innovation (UKRI) Centre for Doctoral Training in Interactive Artificial Intelligence Award (EP/S022937/1). We are greatly thankful to Behrad Koohy for inspiring an earlier version of this work, and to Sennay Ghebreab for providing support for this collaboration.

References

1. De vereniging van schoolbesturen in het amsterdamse voortgezet onderwijs. https://www.verenigingosvo.nl/ (2022), accessed: 2022-08-18
2. Abdulkadiroglu, A., Pathak, P.A., Roth, A.E., Sönmez, T.: Changing the boston school choice mechanism (2006)
3. Abdulkadiroğlu, A., Sönmez, T.: Random serial dictatorship and the core from random endowments in house allocation problems. Econometrica **66**(3), 689–701 (1998)
4. Abdulkadiroğlu, A., Sönmez, T.: School choice: A mechanism design approach. American economic review **93**(3), 729–747 (2003)
5. Agarwal, N., Somaini, P.: Revealed preference analysis of school choice models. Annual Review of Economics **12**(1), 471–501 (2020)
6. Che, Y.K., Grenet, J., He, Y.: Allocating students to schools: theory, methods, and empirical insights. Handbook of the Economics of Matching **2**, 307–407 (2025)
7. Díaz, D.A., Jiménez, A.M., Larroulet, C.: An agent-based model of school choice with information asymmetries. Journal of Simulation **15**(1-2), 130–147 (2021)
8. Erdil, A., Ergin, H.: What's the matter with tie-breaking? improving efficiency in school choice. American Economic Review **98**(3), 669–689 (2008)
9. Gale, D., Shapley, L.S.: College admissions and the stability of marriage. The American mathematical monthly **69**(1), 9–15 (1962)
10. Hermanussen, L., Groot Beumer, T., Grond, A.: Evaluatie van het plaatsingssysteem (2019), https://www.verenigingosvo.nl/wp-content/uploads/2019/09/190904-Bijl.-2-Eindrapport-evaluatie-Dialogic.pdf
11. Riehm, T., van de Scheur, D.: Unintended consequences: Navigating the complexity of school allocation in amsterdam (Sep 2023), https://www.tws-partners.com/2023/09/15/unintended-consequences-navigating-the-complexity-of-school-allocation-in-amsterdam/
12. Roth, A.: Deferred acceptance algorithms: History, theory, practice, and open questions. international Journal of game Theory **36**(3), 537–569 (2008)
13. Skrepr: Procesomschrijving Centrale Loting en Matching Amsterdam. Tech. rep., De vereniging van schoolbesturen in het Amsterdamse voortgezet onderwijs (OSVO) (2026), https://verenigingosvo.nl/wp-content/uploads/2025/12/Bijlage-2-Procesomschrijving-Centrale-Loting-en-Matching-OSVO.pdf
14. Tasnim, M., Verhagen, P., Blanke, T., Acar, E., Ghebreab, S.: Modeling strategic risk in school choice: A case for transparent design. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 8, pp. 2470–2479 (2025)
15. Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L.S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., et al.: Pettingzoo: Gym for multi-agent reinforcement learning. Advances in Neural Information Processing Systems **34**, 15032–15043 (2021)

A RSD with Placement Guarantee Algorithm

Algorithm 1 gives a step-by-step description of the allocation procedure used in the model and simulations. The procedure first runs Random Serial Dictatorship (RSD) [3] using the submitted lists and the original school capacities. Students who cannot be assigned to any school on their submitted list in this main round remain unassigned. Among these students, only those who submitted a list of length at least l are eligible for the placement guarantee. The guarantee stage is then modelled as an ex-post capacity expansion, where each school s receives $\lfloor \gamma c_s \rfloor$ additional seats. If the number of guarantee-eligible students exceeds the total number of additional seats, the guarantee becomes infeasible.

Algorithm 1 RSD with placement guarantee

Require: Students N, schools S, capacities c_s, reported lists L_i, guarantee threshold l, expansion rate γ

Draw a random permutation π of students
Initialise remaining capacities $q_s \leftarrow c_s$ for all $s \in S$
for student i in order π **do**
    Assign i to the highest-ranked school in L_i with $q_s > 0$, if one exists
    If assigned to school s, set $q_s \leftarrow q_s - 1$; otherwise leave i unassigned
end for
Let $U = \{i : i \text{ is unassigned and } |L_i| \geq l\}$
Expand capacities by $\tilde{c}_s = c_s + \lfloor \gamma c_s \rfloor$ for all $s \in S$
Reconsider students in U using expanded capacity in order of π
If $|U| > \sum_{s \in S} \lfloor \gamma c_s \rfloor$, the guarantee is infeasible

B Strategic Reporting under Placement Guarantee

We focus on the incentives of a single student i, taking the behaviour of all other students as fixed. Student i has sincere preferences over at most $k < l$ schools, $r_1 \succ_i r_2 \succ_i \cdots \succ_i r_k$, and is indifferent between all other schools except for their effect on eligibility for the placement guarantee. Let $u_i(s)$ denote the utility of i for school s, with $u_i(r_j) > 0$ for $1 \leq j \leq k$ and $u_i(s) = 0$ for any school that i considers unacceptable.

Let $P \subseteq S$ be the set of popular schools introduced in Section 2, with $|P| \geq l - k$. The *top-k truthful* strategy for student i is the reported list $L_i^{\mathrm{TP}} = [r_1, \ldots, r_k, p_1, \ldots, p_{l-k}]$, where $(p_1, \ldots, p_{l-k})$ is any ordering of $l - k$ distinct schools from P. Thus i reports the sincere top k schools and uses popular schools to fill the remaining positions up to the guarantee threshold l.

We restrict attention to reported lists that agree with the sincere ranking in the top k positions and differ only in the tail. Formally, let $\mathcal{L}_i$ be the set of lists $L_i = [r_1, \ldots, r_k, t_{k+1}, \ldots, t_l]$, where $t_{k+1}, \ldots, t_l$ are distinct schools in $S \setminus \{r_1, \ldots, r_k\}$. By construction, any school in positions $k+1, \ldots, l$ yields zero direct utility for i.

Assumption 1 *For student i, the expected utility from the placement guarantee depends only on the top k positions of L_i and is the same for all lists $L_i \in \mathcal{L}_i$. Moreover, schools in P are sufficiently oversubscribed that the probability of i being assigned to any $p \in P$ in the main round is negligible.*

For a reported list $L_i = [s_1, \ldots, s_l]$, let $p_j(L_i)$ denote the probability that student i is assigned to s_j in the main RSD round, and let $P_{\mathrm{MR}}(L_i) = \sum_{j=1}^{l} p_j(L_i)$ be the probability that i is assigned to some school on the list in the main round. Let $u_i^{\mathrm{PG}}(L_i)$ denote the expected utility from the placement guarantee stage, conditional on being unassigned and guarantee-eligible. The expected utility of i under L_i can be written as

$$\mathbb{E}[U_i(L_i)] = \sum_{j=1}^{l} p_j(L_i)\, u_i(s_j) + \big(1 - P_{\mathrm{MR}}(L_i)\big)\, u_i^{\mathrm{PG}}(L_i).$$

For lists $L_i \in \mathcal{L}_i$, the top k positions coincide with $[r_1, \ldots, r_k]$, and all tail schools $t_{k+1}, \ldots, t_l$ yield zero direct utility. Hence the first term reduces to $\sum_{j=1}^{k} p_j(L_i)\, u_i(r_j)$. Under Assumption 1, $u_i^{\mathrm{PG}}(L_i)$ is the same for all $L_i \in \mathcal{L}_i$, so the tail affects the second term only through $P_{\mathrm{MR}}(L_i)$.

Proposition 1. *Consider a student i with sincere preferences over at most $k < l$ schools. Under Assumption 1, any list in $\mathcal{L}_i$ that uses only schools from P in positions $k+1, \ldots, l$ weakly dominates any list in $\mathcal{L}_i$ that uses at least one other school in the tail. In particular, the top-k truthful strategy L_i^{TP} is a best response within $\mathcal{L}_i$.*

Proof (Proof sketch). Replacing a tail school that may assign i in the main round by a popular school $p \in P$ with negligible assignment probability leaves the distribution over $\{r_1, \ldots, r_k\}$ unchanged and weakly reduces $P_{\mathrm{MR}}(L_i)$. Since $u_i^{\mathrm{PG}}(L_i)$ is unaffected within $\mathcal{L}_i$, this weakly increases the weight on the guarantee term while preserving the first term, and thus weakly increases $\mathbb{E}[U_i(L_i)]$. Iterating this replacement over all tail positions yields the claim.

3.3 The Role of Regulatory Institutions in Strategic Lending

The Role of Regulatory Institutions in Strategic Lending

Mayesha Tasnim, Marta C. Couto, and Giovanni Sileno

Socially Intelligent Artificial Systems, University of Amsterdam
{m.tasnim, m.gomesdacunhacouto, g.sileno}@uva.nl

Abstract. This paper studies how a regulatory institution may affect outcomes in a minimal strategic lending model with heterogeneous borrowers. We consider a one-shot setting with three classes of agents: a profit-maximizing lender that applies a threshold rule to a reported borrower score, a population of borrowers that can strategically adjust this score at a cost, and a regulator that shapes the lender's incentives according to a given social policy. Each borrower has a true creditworthiness level and an adjustment cost parameter; a fixed threshold on true creditworthiness separates *good* (creditworthy) from *bad* borrowers, following the good/bad distinction in strategic classification models. Borrowers choose score adjustments to maximize their own payoff, trading off the benefit of receiving a loan against their individual adjustment costs and, for good borrowers, the harm of being denied credit. The lender's profit depends on the mix of accepted good and bad borrowers, while a regulator introduces a penalty on the number of good borrowers who are denied credit and controls the strength of this penalty. We compare the resulting regulated regime to an unregulated baseline and use simulation experiments that vary regulatory strength and heterogeneity in adjustment costs to assess how regulation may shift credit access and lender profit in this stylized setting.

Keywords: Strategic classification · Incentive design · Credit lending

1 Introduction

Many institutional decisions in credit, housing, or hiring are implemented as simple threshold rules on observable scores, often produced by predictive models [7, 2, 4]. In credit settings, such scores may include observable variables such as reported income, debt-to-income ratio, recent repayment history, or a composite credit score, and a lender may approve an application only if this reported score exceeds a threshold. A growing literature on strategic classification highlights that these scores are not fixed: individuals can adapt their observable features in response to a decision rule, often at a personal cost [5, 8, 11, 1]. In lending, such adaptation can be interpreted as paying effort to improve or otherwise adjust (i.e. manipulating perception of) a profile to clear a credit threshold. Critically, these effort costs are heterogeneous: some borrowers can adjust their score cheaply,

while others face much higher barriers, which can create systematic disparities in who is able to respond to a model [12, 6]. As a result, even a formally neutral threshold rule can induce unequal strategic burdens and unequal access to loans. Our focus differs from most existing strategic-classification work, which studies strategic populations responding to a classifier, by explicitly modelling a regulator as an additional institutional agent that intervenes in the lender's objective.

In this paper we focus on the institutional role of regulation in such a setting. We develop a minimal model of strategic lending with regulation with three types of agents: a profit-seeking lender, a population of borrowers with heterogeneous adjustment costs, and a regulator. Each borrower has a true creditworthiness level and can pay a cost (in this model quadratic) to increase the score that the lender observes, similar in spirit to effort-based models of strategic classification [5, 8]. A fixed threshold on true creditworthiness separates good (creditworthy) from bad borrowers and is used only to identify when denial of credit is socially costly, following the good/bad distinction in prior work [12, 6, 3]. The lender applies a threshold to reported scores and earns positive expected profit on loans to good borrowers and negative expected profit on loans to bad borrowers.

The regulator is concerned with exclusions of creditworthy borrowers and with the broader sociotechnical harms of such exclusions [14, 9]. Rather than directly setting the lender's threshold, it introduces a penalty on the number of good borrowers who are denied and controls the strength of this penalty through a parameter λ. The lender then chooses its threshold to trade off profit against the penalised harm, yielding an equilibrium that depends on regulatory strength. Our focus is complementary to recent work on performative prediction, which studies how predictive models and outcomes co-evolve [13, 10], by explicitly introducing a regulatory agent that penalises exclusions within a stylised strategic lending environment. Our model can be viewed as a minimal, one-shot variant of the collective strategic-classification dynamics studied by [3], and our experiments illustrate how even a simple regulatory penalty can shift the lender from a high-profit, high-exclusion regime to a more inclusive one. Specifically, we study how varying λ changes credit access and lender profit, and the distribution of adjustment costs borne by borrowers, with particular attention to disadvantaged groups who pay a higher cost for strategic adjustments. This one-shot formulation is intended as a minimal static benchmark: it isolates the immediate threshold response of the lender and borrowers to a given regulatory penalty.

2 Model

We consider a one-shot interaction between three types of agents: a lender, a population of borrowers, and a regulator. Decisions are made via a simple threshold rule on a scalar score, following the strategic classification setup of [3].

Borrowers. There are N borrowers indexed by $i \in \{1, \ldots, N\}$. Each borrower has a *true creditworthiness* $z_i \in \mathbb{R}$ and an *adjustment cost* parameter $k_i > 0$. A fixed

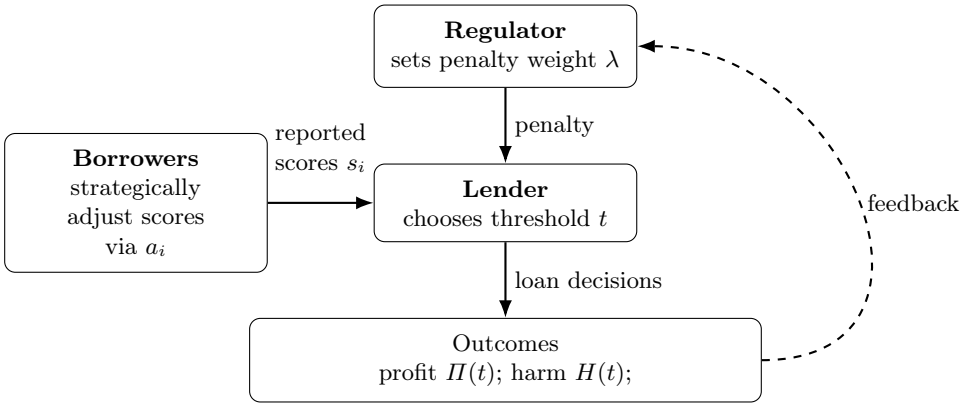


Fig. 1: Conceptual flow of the regulated strategic lending model: the regulator sets a penalty weight on excluding creditworthy borrowers, the lender chooses a threshold on reported scores, borrowers strategically adjust their scores, and outcomes determine profit and credit access across heterogeneous borrowers.

threshold θ on true creditworthiness is used to distinguish *good* (creditworthy) from *bad* borrowers:

$$\text{good if } z_i \geq \theta, \qquad \text{bad if } z_i < \theta,$$

Borrowers can strategically adjust the score used by the lender through a single action $a_i \geq 0$. The reported score is

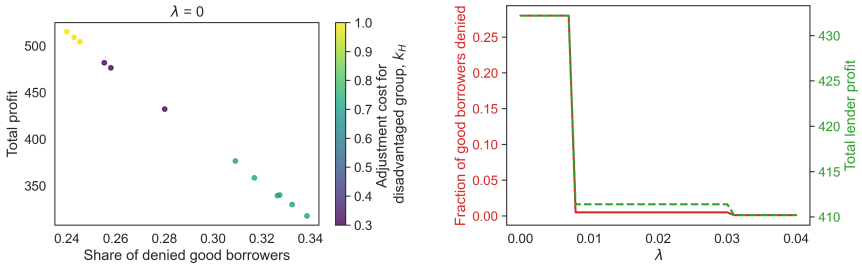
$$s_i = z_i + a_i,$$

and making a larger adjustment is more costly. We use a quadratic cost $c_i(a_i) = k_i a_i^2$ and capture heterogeneity in strategic capacity by assigning different k_i values to different groups (e.g., a low-cost group and a disadvantaged high-cost group). Receiving a loan yields benefit $b > 0$. If a good borrower is rejected, they incur an additional loss $h > 0$ representing the opportunity cost of exclusion; bad borrowers do not incur this loss when rejected. Given a threshold t, each borrower chooses a_i to maximize their expected payoff (benefit minus cost and possible exclusion loss).

Lender. The lender applies a threshold policy t (distinct from θ) on the reported score: borrower i is accepted if $s_i \geq t$ and rejected otherwise. Loans to good borrowers are profitable on average, while loans to bad borrowers are unprofitable. We summarize this by two constants $\pi^G > 0$ and $\pi^B < 0$, denoting expected profit from an accepted good or bad borrower, respectively, and set profit to zero when a borrower is rejected. Total expected profit at threshold t is then

$$\Pi(t) = \pi^G \cdot \#\{i : z_i \geq \theta,\ s_i \geq t\} + \pi^B \cdot \#\{i : z_i < \theta,\ s_i \geq t\}, \tag{1}$$

where the counts are determined by borrowers' best responses $a_i(t)$.



(a) Profit and Costs vs. share of good borrowers who are denied

(b) Effect of λ on harm and profit.

Fig. 2: Simulation results for the regulated strategic lending model. The left panel shows unregulated regimes ($\lambda = 0$): each point shows the total profit and the fraction of good borrowers who are denied a loan when the lender selects the most profiting threshold; colour encodes the adjustment parameter k_H for the disadvantaged group. The right panel shows the fraction of good borrowers denied (left axis) and total profit (right axis) as functions of the penalty weight λ, highlighting a sharp drop in harm at modest profit loss.

Regulator. The regulator is concerned with the exclusion of creditworthy borrowers. For a given threshold t, we define the harm

$$H(t) = \#\{i : z_i \geq \theta,\ s_i < t\},$$

i.e., the number of good borrowers who are denied a loan. Regulation acts by introducing a penalty on this harm into the lender's objective. For a penalty weight $\lambda \geq 0$, the lender solves

$$t(\lambda) \in \arg\max_t \big(\Pi(t) - \lambda H(t)\big). \tag{2}$$

When $\lambda = 0$, the lender purely maximizes profit; as λ increases, thresholds that exclude many good borrowers become less attractive.

Social interaction. The interaction amongst borrowers, lender and regulator proceeds in four steps:

(i) the regulator fixes a penalty weight λ
(ii) the lender chooses a threshold $t(\lambda)$
(iii) borrowers best-respond by choosing $a_i(t(\lambda))$
(iv) outcomes such as the lender profit $\Pi(t(\lambda))$ and harm $H(t(\lambda))$ are determined

With this set-up, by varying λ and the heterogeneity in k_i we can study how regulation affects credit access, strategic burden, and inequality between low-cost and high-cost borrowers.

3 Experiments

We run two simple simulations. In all runs we use $N = 10{,}000$ borrowers, draw $z_i \sim \mathcal{N}(0, 1)$, set the good/bad cutoff at $\theta = 0$, fix $\pi_G = 0.2$, and let borrowers best-respond to each threshold t. For any (t, λ), we compute profit $\Pi(t)$ and harm $H(t)$, which is measured by the number of good borrowers that were denied a loan. The lender chooses t^* via a grid search over t.[1]

First, we study the unregulated case $\lambda = 0$ across a grid of environments, varying the share of low-cost borrowers $p_L \in \{0.3, 0.5\}$, borrower benefit $b \in \{0.8, 1.0\}$, harm parameter $h \in \{0.1, 0.3, 0.5\}$, high-cost adjustment cost $k_H \in \{0.3, 0.7, 1.0\}$ (with $k_L = 0.1$ fixed), and profit on bad borrowers $\pi_B \in \{-0.4, -0.6\}$. For each setting we find $t^*(0)$ (the best threshold the lender can set) and record total profit, the fraction of good borrowers denied, and group-specific acceptance rates. Figure 2a plots the fraction of good borrowers denied versus $\Pi(t^*(0))$, with colour encoding k_H. We observe that some of the highest-profit regimes still exclude a large share (around 25–35%) of good borrowers, and that these high-profit, high-exclusion points are associated with particular values of k_H, but the pattern is clearly non-linear; understanding more precisely how k_H shapes this profit–exclusion frontier is an interesting direction for future work.

Second, we fix a representative high-harm regime from this grid with $p_L = 0.3$, $b = 1.0$, $h = 0.1$, $k_L = 0.1$, $k_H = 0.3$, and $\pi_B = -0.6$, and vary λ on a fine grid. Next, for each λ we compute $t^*(\lambda)$, $\Pi(t^*(\lambda))$, and the fraction of good borrowers denied. Figure 2b shows that the harm is high and flat for small λ, until a critical value $\lambda^\star$ where the lender switches to a lower threshold: the fraction of good borrowers denied drops sharply (from about 28% to below 1%), while profit decreases less dramatically (from about 432 to 412).

4 Discussion and Future Work

The present paper is best read as a proof-of-concept showing that even a minimal regulatory intervention can alter equilibrium lending behaviour in a strategic environment. Our initial experiments highlight two simple observations. First, in the absence of regulation, profit-maximizing thresholds can coincide with substantial exclusion of creditworthy borrowers, especially when the disadvantaged group faces high adjustment costs. Second, it suggests that even a very simple regulatory lever can induce a discrete shift in the lender's best response: a small increase in the penalty weight sharply reduces harm while reducing lender profit by $\approx 5\%$. In this stylized setting, modest regulatory pressure can influence lender incentives without fully undermining profitability.

The present model is intentionally minimal. It is one-shot, with a single lender, binary good/bad types, and a linear penalty on exclusion, and thus abstracts away from many institutional and market details of real-world credit

[1] The code for the experiments are available at https://github.com/m-tasnim/regulated-strategic-lending.git

systems [4]. In particular, the binary good/bad distinction compresses a continuous repayment-risk distribution into a hard cutoff and therefore abstracts away from heterogeneity near the boundary of creditworthiness. Despite this simplification, the model highlights how even a basic regulatory penalty can shift incentives in the presence of heterogeneous adjustment costs. Because the model is one-shot, it should be interpreted as identifying short-run comparative statics rather than long-run institutional dynamics. Future work could extend it along several dimensions: richer borrower action spaces (e.g., separating *faking* creditworthiness from *improving*), alternative welfare and penalty functions, multiple competing lenders, and dynamic settings in which both lenders and regulators learn or adapt over time. Calibrating the model to empirical credit data or to more detailed strategic-classification models would also be a natural next step.

A central simplifying assumption of the model is that the regulator can perfectly identify all good borrowers who were denied credit. This likely overstates the effectiveness of regulatory penalties. In practice, such cases would only be partially observable, for example through borrower complaints backed by documentation, supervisory audits of lender decisions, or ex-post outcome monitoring by an oversight body. Extending the model to partial or noisy observability is therefore one of the most important directions for future work.

Acknowledgements

Work partly funded by NWO for the HUMAINER AI project (KIVI.2019.006).

References

1. Barsotti, F., Koçer, R.G., Santos, F.P.: Transparency, detection and imitation in strategic classification. In: IJCAI. pp. 67–73 (2022)
2. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery **21**(2), 277–292 (2010)
3. Couto, M.C., Barsotti, F., Santos, F.P.: Collective dynamics of strategic classification (2025), https://arxiv.org/abs/2508.09340
4. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A.: Predictably unequal? the effects of machine learning on credit markets. The Journal of Finance **77**(1), 5–47 (2022)
5. Hardt, M., Megiddo, N., Papadimitriou, C., Wootters, M.: Strategic classification. In: Proceedings of the 2016 ACM conference on innovations in theoretical computer science. pp. 111–122 (2016)
6. Hu, L., Immorlica, N., Vaughan, J.W.: The disparate effects of strategic manipulation. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 259–268 (2019)
7. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions. The quarterly journal of economics **133**(1), 237–293 (2018)
8. Kleinberg, J., Raghavan, M.: How do classifiers induce agents to invest effort strategically? ACM Transactions on Economics and Computation (TEAC) **8**(4), 1–23 (2020)

9. Kroll, J.A.: Accountable algorithms. Princeton University (2015)
10. Mendler-Dünner, C., Perdomo, J., Zrnic, T., Hardt, M.: Stochastic optimization for performative prediction. Advances in Neural Information Processing Systems **33**, 4929–4939 (2020)
11. Miller, J., Milli, S., Hardt, M.: Strategic classification is causal modeling in disguise. In: International Conference on Machine Learning. pp. 6917–6926. PMLR (2020)
12. Milli, S., Miller, J., Dragan, A.D., Hardt, M.: The social cost of strategic classification. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 230–239 (2019)
13. Perdomo, J., Zrnic, T., Mendler-Dünner, C., Hardt, M.: Performative prediction. In: International Conference on Machine Learning. pp. 7599–7609. PMLR (2020)
14. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 59–68 (2019)

4 Multiagent Learning, Cooperation, and Coordination

4.1 Resolving Complex Social Dilemmas by Aligning Preferences with Counterfactual Regret

Resolving Complex Social Dilemmas by Aligning Preferences with Counterfactual Regret

Shuqing Shi[1], Yudi Zhang[2], Joel Z. Leibo[3], and Yali Du[1]

[1] King's College London
[2] Eindhoven University of Technology
[3] Google Deepmind

Abstract. Social dilemmas are situations where gains from cooperation are possible but misaligned incentives make it hard to stabilize prosocial joint behavior. In spatiotemporally complex social dilemmas, barriers from misaligned incentives interact with obstacles from spatiotemporal complexity. We propose a multi-agent reinforcement learning algorithm that finds cooperative resolutions by having agents minimize counterfactual regret—the gap between optimal prosocial behavior and current behavior. This approach disentangles the causes of selfish reward from prosocial reward. Empirically, our method outperforms multiple baselines in several complex social dilemma environments.

1 Introduction

Individuals often have desires that do not align with their group's objectives. When such scenarios contain spatial and temporal complexity they are called sequential social dilemmas (SSDs) (Leibo et al., 2017). SSDs are challenging because spatial and temporal complexity interacts with strategic complexity from misaligned incentives. For example, in Cleanup (Hughes et al., 2018), players collect apples whose growth depends on cleaning a polluted river—a prosocial but costly action, creating a division of labor where cleaners sacrifice foraging time for others' benefit (Yaman et al., 2023).

Many works promote cooperation in SSDs: LOLA (Foerster et al., 2017) models opponents' behaviors; Jaques et al. (2019) investigate causal relationships between agents' actions; intrinsic motivation approaches (Hughes et al., 2018; McKee et al., 2020; Lupu & Precup, 2020; Kwon et al., 2023) encourage agents to maximize others' welfare; D3C (Gemp et al., 2020) aligns incentives automatically. However, these approaches fail to capture the reward generation process, potentially leading to spurious predictions of true team incentives. Prior centralized training methods (Foerster et al., 2016, 2018) ensure coordination but do not address the entanglement of agents' policies.

We propose using counterfactual regret to align incentives. In SSDs, agents may be rewarded for selfish behaviors when others cooperate, causing entanglement that biases contribution estimation. We utilize a causal model to capture the reward generation process and define counterfactual regret as the difference

between the maximum counterfactual reward for other agents and their actual reward. Minimizing this regret guides agents to consider others' rewards. Our contributions are: (1) a generative model capturing reward generation in SSDs for counterfactual reasoning; (2) counterfactual regret inference to construct intrinsic rewards aligning agents with social incentives; (3) comprehensive evaluation on four SSD tasks demonstrating superior performance. Our method adopts Centralized Training with Decentralized Execution (CTDE), requiring joint observations during training while enabling fully decentralized execution.

2 Methodology

2.1 Preliminaries

Partially Observable Markov Game (POMG) is defined by $\langle N, S, O, T, A, R\rangle$. At each timestep t, agent $i \in N$ chooses action $a_t^i \in A$. The joint action $\boldsymbol{a}_t = [a_t^1, \cdots, a_t^N]$ produces a transition $T(s_{t+1}|\boldsymbol{a}_t, s_t)$. Agent i observes o_t^i and maximizes $R^i = \sum_{t=0}^{\infty} \gamma^t r_t^i$, where reward $r^i(\boldsymbol{a}_t, s_t)$ depends on other agents' actions.

Counterfactual Reasoning refers to reasoning about reward changes under different actions. Under Markovian graphs (Pearl, 2010, 2009), conditioning on state $\boldsymbol{s}$: $P(\boldsymbol{r}^{\text{cf}}|do(\boldsymbol{a} = \boldsymbol{a}^{\text{cf}}), \boldsymbol{s}) = P(\boldsymbol{r}^{\text{cf}} \mid \boldsymbol{a}^{\text{cf}}, \boldsymbol{s})$. In our POMG setting, conditioning on $\boldsymbol{s}$ and $\boldsymbol{a}^{-i}$ d-separates $\boldsymbol{r}^{-i}$ from $\boldsymbol{a}^i$, allowing counterfactual reward estimation.

2.2 Overview

Each agent i has a **generative model** Φ_{m}^i predicting rewards given joint observation $\boldsymbol{o}_t$ and action $\boldsymbol{a}_t$, and a **policy model** Φ_π^i mapping o_t^i to a_t^i. The overall objective is:

$$L^i(\Phi_{\text{m}}^i, \Phi_\pi^i) = L_{\text{m}}^i(\Phi_{\text{m}}^i) + L_\pi^i(\Phi_\pi^i), \tag{1}$$

where L_{m}^i is defined in Eq. 5 and L_π^i in Eq. 6. The intrinsic reward is the negative of counterfactual regret, promoting cooperative behaviors. Figure 1 depicts the framework.

2.3 Causal Modeling

We model the environment via a Dynamic Bayesian Network $\mathcal{G}$ over $[\boldsymbol{s}_t, \boldsymbol{o}_t, \boldsymbol{a}_t, \boldsymbol{r}_t]_{t=1}^T$:

$$\begin{cases} \boldsymbol{s}_{t+1} = f(\boldsymbol{s}_t, \boldsymbol{a}_t, \epsilon_{\boldsymbol{s},t}) & \text{(transition)} \\ \boldsymbol{r}_t^i = g^i(\boldsymbol{s}_t, \boldsymbol{a}_t, \epsilon_{r,i,t}) & \text{(reward)} \\ \boldsymbol{o}_t^i = h^i(\boldsymbol{s}_t, \boldsymbol{a}_t, \epsilon_{o,i,t}) & \text{(observation)} \end{cases} \tag{2}$$

where $\epsilon_\cdot$ denotes i.i.d. noise and $\mathcal{G}$ is time-invariant with no unobserved confounders (Huang et al., 2021).

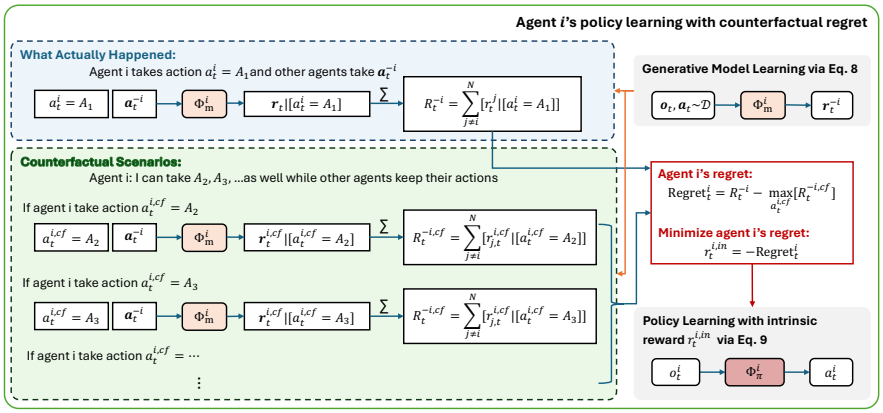


Fig. 1: Training and inference of agent i. Blue: actual reward generation; green: counterfactual-reward generation; red: regret calculation and intrinsic-reward construction; gray: learning processes (Generative Model via Eq. 5, Policy via Eq. 6).

Proposition 1 (Blockwise identifiability; proof in Appendix A). *Under Assumption 1, for each reward r^i, the learned latent block $\hat{s}^{r^i}$ depends only on the true reward-relevant block s^{r^i} up to an invertible reparameterization $\hat{s}^{r^i} = H_i(s^{r^i})$, with no leakage from reward-irrelevant coordinates.*

This justifies learning a mapping from observations to rewards via neural networks. Under d-separation conditions, the shaped game preserves equilibria of the original game (see Appendix A).

2.4 Counterfactual Regret Generation

Counterfactual individual rewards. We ask: how much would other agents earn if agent i takes $a_t^{i,\mathrm{cf}}$ instead of a_t^i? Using generative model $\Phi_{\mathrm{m}}^i : \mathcal{O}^N \times \mathcal{A}^N \to \mathbb{R}^N$:

$$\boldsymbol{r}_t^{i,\mathrm{cf}} = \Phi_{\mathrm{m}}^i(\boldsymbol{o}_t, a_t^{i,\mathrm{cf}}, \boldsymbol{a}_t^{-i}), \tag{3}$$

where $\boldsymbol{a}_t^{-i}$ denotes actions of agents excluding i. The collective counterfactual reward is $R_t^{-i,\mathrm{cf}} = \sum_{j\neq i}^{N} \Phi_{\mathrm{m},j}^i(\boldsymbol{o}_t, a_t^{i,\mathrm{cf}}, \boldsymbol{a}_t^{-i})$.

Counterfactual Regret. For agent i:

$$\mathrm{Regret}_t^i = \max_{a_t^{i,\mathrm{cf}}}[R_t^{-i,\mathrm{cf}}(\boldsymbol{o}_t, a_t^{i,\mathrm{cf}}, a_t^{-i})] - R_t^{-i}(\boldsymbol{o}_t, a_t^i, a_t^{-i}), \tag{4}$$

where $R_t^{-i} = \sum_{j\neq i}^{N} \Phi_{\mathrm{m},j}^i(\boldsymbol{o}_t, a_t^i, \boldsymbol{a}_t^{-i})$. We compute the max by enumeration over $\mathcal{A}^i$ (typically $|\mathcal{A}^i| \leq 9$).

Intrinsic Reward. We set $r_t^{i,\mathrm{in}} = -\mathrm{Regret}_t^i$, yielding shaped reward $\hat{r}_t^i = r_t^{i,\mathrm{ex}} + \alpha r_t^{i,\mathrm{in}}$, where α controls the weight on others' welfare.

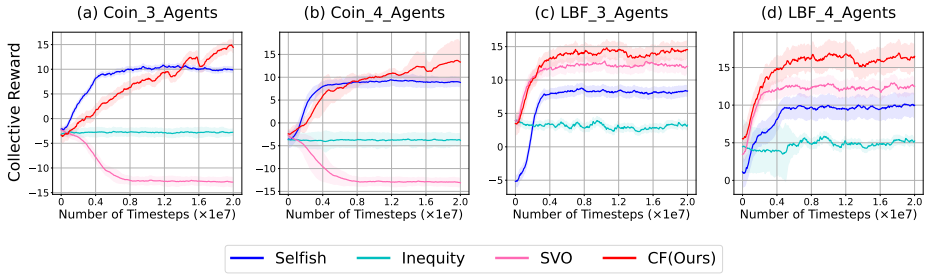


Fig. 2: Collective reward on *Coin* and *LBF* (3 and 4 agents), 5 runs. Shaded: std. dev. Training: 2×10^7 steps.

2.5 Learning Objectives

Generative Model. For each agent i:

$$L_m^i = \mathbb{E}_{\boldsymbol{o}_t, \boldsymbol{a}_t, \boldsymbol{r}_t \sim D} \left[||\Phi_{\mathrm{m}}^i(\boldsymbol{o}_t, \boldsymbol{a}_t) - \boldsymbol{r}_t^{\mathrm{ex}}||^2 \right]. \tag{5}$$

Policy Learning. Using PPO (Schulman et al., 2015):

$$L_\pi^i = \hat{\mathbb{E}}_t \left[\min(\hat{r}_t^i(\theta)\hat{A}_t^i, \mathrm{clip}(\hat{r}_t^i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t^i) \right], \quad \text{where } \hat{A}_t^i = Q_t^i(o_t^i, a_t^i) - V_t^i(o_t^i). \tag{6}$$

During training, intrinsic rewards and Φ_m use joint observations (centralized). At execution, each Φ_π^i uses only local o_t^i (decentralized). Both networks use GRU layers for partial observability. The full algorithm pseudocode is in Appendix B.3.

3 Experiments

3.1 Setup

Environments. We evaluate on four SSDs: *Coin* (Lerer & Peysakhovich, 2017) (type-matching coin collection), *LBF* (Christianos et al., 2020) (cooperative foraging with doubled rewards), *Cleanup* (Hughes et al., 2018) and *Common_Harvest* (Perolat et al., 2017), following Jaques et al. (2019). We test scalability with 3–7 agent variants (Appendix B.1).

Baselines. PPO (Schulman et al., 2015) (Selfish), inequity aversion (Hughes et al., 2018) (Inequity), and SVO (McKee et al., 2020). Details in Appendix B.3.

3.2 Results

Figures 2 and 3 show that our method (CF) consistently outperforms all baselines across all eight scenarios. In *Coin* (Fig. 2a-b), CF maintains a significant advantage throughout training. *Coin_4_Agents* includes an adversarial agent (receives +1

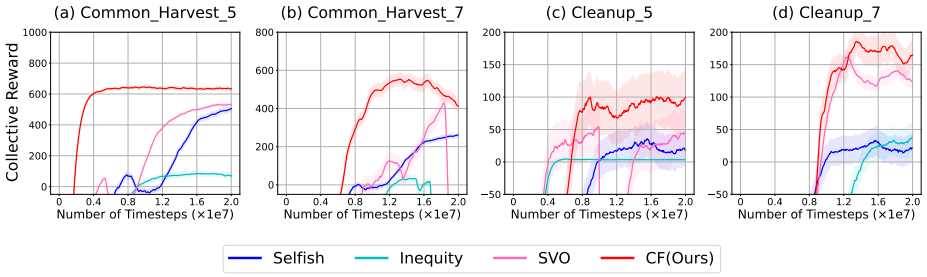


Fig. 3: Collective reward on *Common_Harvest* and *Cleanup* (5 and 7 agents), 5 runs. Shaded: std. dev. Training: 2×10^7 steps.

per coin, victim receives -2); CF handles this chaos and still outperforms baselines. In *LBF* (Fig. 2c-d), CF shows remarkable stability, with pronounced gaps in the 4-agent setting.

For *Common_Harvest* (Fig. 3a-b), a tragedy-of-the-commons scenario, CF significantly outperforms all baselines with strong scalability from 5 to 7 agents. In *Cleanup* (Fig. 3c-d), a public goods game requiring self-sacrifice, CF achieves the highest collective reward with consistent improvement. Ablation studies (Appendix) show the optimal α scales as $(N-1)$: $\alpha = 2$ for 3-agent and $\alpha = 5$ for 4-agent settings, where each agent equally weights each other individual's welfare relative to its own.

4 Conclusion

We propose a multi-agent reinforcement learning algorithm that addresses social dilemmas by encouraging agents to minimize counterfactual regret—the gap between optimal prosocial behavior and current behavior. This enables agents to balance self-interest with cooperation by disentangling selfish and prosocial rewards. Empirical evaluations demonstrate consistent improvements over baselines across complex social dilemma environments. Future work will focus on developing robust strategies against exploitation by defectors.

Bibliography

Christianos, F., Schäfer, L., and Albrecht, S. V. Shared experience actor-critic for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Foerster, J., Assael, I. A., De Freitas, N., and Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

Gemp, I., McKee, K. R., Everett, R., Duéñez-Guzmán, E. A., Bachrach, Y., Balduzzi, D., and Tacchetti, A. D3c: Reducing the price of anarchy in multi-agent learning. *arXiv preprint arXiv:2010.00575*, 2020.

Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., and Wu, Y. Stable baselines. `https://github.com/hill-a/stable-baselines`, 2018.

Huang, B., Feng, F., Lu, C., Magliacane, S., and Zhang, K. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*, 2021.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Kwon, M., Agapiou, J. P., Duéñez-Guzmán, E. A., Elie, R., Piliouras, G., Bullard, K., and Gemp, I. Auto-aligning multiagent incentives with global objectives. In *ICML Workshop on Localized Learning (LLW)*, 2023.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473, 2017.

Lerer, A. and Peysakhovich, A. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

Lupu, A. and Precup, D. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on autonomous agents and multiagent systems*, pp. 789–797, 2020.

McKee, K. R., Gemp, I., McWilliams, B., Duèñez-Guzmán, E. A., Hughes, E., and Leibo, J. Z. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 869–877, 2020.

Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96, 2009.

Pearl, J. Causal inference. In Guyon, I., Janzing, D., and Schölkopf, B. (eds.), *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Proceedings of Machine Learning Research*, pp. 39–58, Whistler, Canada, 12 Dec 2010. PMLR. URL `https://proceedings.mlr.press/v6/pearl10a.html`.

Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, 30, 2017.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Yaman, A., Leibo, J. Z., Iacca, G., and Wan Lee, S. The emergence of division of labour through decentralized social sanctioning. *Proceedings of the Royal Society B*, 290(2009):20231716, 2023.

A Details on proofs

Given the joint observations $\boldsymbol{o}_t^i,\ \forall i \in [1, \dots N]$, joint action $\boldsymbol{a}_t$, we prove that, reward-relevant $\boldsymbol{s}_t^{r^i}$ is identifiable, as well as the unknown functions. Our generative model in Eq. 2, denoted by a Dynamic Bayesian Network (DBN) $\mathcal{G}$, is constructed over the variables $\{\{\boldsymbol{s}_t^k\}^{|s|}, \{\boldsymbol{o}_t^i, \boldsymbol{a}_t^i, \boldsymbol{r}_t^i\}^N\}^T$ in Partially Observable Markov Game.

According to (Pearl, 2010), we can do counterfactual reasoning if we know all the causal parents of the variable $\boldsymbol{r}_t^i$. Therefore, the goal is to show that we can identify an agent i's reward-relevant set of state components $\boldsymbol{s}_t^{r^i}$ which have a direct path to the individual rewards $\boldsymbol{r}_t^i$.

We show how we identify the latent state $\boldsymbol{s}_t^{r^i}$.

We rewrite the reward function and observation function in Eq 2 as,

$$[\boldsymbol{o}_t, \boldsymbol{r}_t^i] = g(\boldsymbol{s}_t^{r^i}, \bar{\boldsymbol{s}}_t^{r^i}, \boldsymbol{a}_t, \epsilon_t),$$

where $\boldsymbol{o}_t$ is the observed variable, and each $\boldsymbol{r}_t^i$ is defined as $\boldsymbol{r}_t^i = g_i(\boldsymbol{s}_t^{r^i}, \boldsymbol{a}_t, \epsilon_{r,i,t})$. $\boldsymbol{s}_t^{r^i}$ is the reward-relevant state components while $\bar{\boldsymbol{s}}_t^{r^i}$ denotes the reward-irrelevant state components.

For simplicity of notation, we omit the time index t below. We denote the dimensions of variables by $d(\cdot)$.

Let $h := \hat{g}^{-1} \circ g$ denote the reparameterization map from true latent variables s to learned latent variables $\hat{s}$. Then we have:

$$\hat{\boldsymbol{s}} = h(\boldsymbol{s}), \quad \text{and equivalently} \quad \boldsymbol{s} = h^{-1}(\hat{\boldsymbol{s}}).$$

Here, $\hat{g} : \mathcal{S} \to \mathcal{X}$ is the estimated generating function, and $g : \mathcal{S} \to \mathcal{X}$ is the true data-generating process. Since both g^{-1} and $\hat{g}$ are assumed to be smooth and invertible, the composite function h and its inverse h^{-1} are also smooth and invertible.

We denote by $\mathcal{F}$ the support of the Jacobian $\boldsymbol{J}_g(\boldsymbol{s})$, by $\hat{\mathcal{F}}$ the support of $\boldsymbol{J}_{\hat{g}}(\hat{\boldsymbol{s}})$, and by $\mathcal{T}$ the support of the transformation $\boldsymbol{T}(\boldsymbol{s})$. We use T to represent a matrix that shares the same sparsity pattern as $\mathcal{T}$.

Standing notation. Let $y := (o^\top, r^\top)^\top \in \mathbb{R}^{d_o+N}$ denote the stacked observations and rewards. The true (unknown) data-generating diffeomorphism is $g : \mathcal{S} \to \mathcal{Y}$, so that $y = g(s)$ for $s \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$. A learned (invertible) model $\hat{g} : \hat{\mathcal{S}} \to \mathcal{Y}$ satisfies $y = \hat{g}(\hat{s})$ for $\hat{s} \in \hat{\mathcal{S}}$.

Reparameterization map: Define $h : \mathcal{S} \to \hat{\mathcal{S}}$ by $h := \hat{g}^{-1} \circ g$, so that $\hat{s} = h(s)$. Equivalently, $h^{-1} : \hat{\mathcal{S}} \to \mathcal{S}$ satisfies $s = h^{-1}(\hat{s})$, and we have the composition $\hat{g} = g \circ h^{-1}$.

Notational convention: Throughout this appendix, we consistently use h to map from true latents to learned latents ($\hat{s} = h(s)$), and h^{-1} for the inverse mapping ($s = h^{-1}(\hat{s})$).

For a matrix M, let

$$\operatorname{supp}(M) \;:=\; \{(i,j) :\; M_{ij} \neq 0\}, \qquad \operatorname{supp}(M)_{i,:} \;:=\; \{j :\; M_{ij} \neq 0\}.$$

Write $\mathcal{F} := \operatorname{supp}(J_g)$, $\hat{\mathcal{F}} := \operatorname{supp}(J_{\hat{g}})$, and $\mathcal{T} := \operatorname{supp}(J_{h^{-1}})$.

Assumption 1 (Invertible generators, row–support, and span richness) *[(i)]*

g and $\hat{g}$ are $\mathcal{C}^1$ diffeomorphisms onto their images. Consequently, $h = g^{-1} \circ \hat{g}$ and h^{-1} are $\mathcal{C}^1$ diffeomorphisms between $\hat{\mathcal{S}}$ and $\mathcal{S}$.

2. *(*Row–support structure of g for rewards*) For each reward coordinate r^i (row index $i \in \{d_o+1, \ldots, d_o+N\}$), there exists an index set $\mathcal{F}_{i,:} \subseteq \{1, \ldots, d_s\}$ (the reward-relevant latent coordinates s^{r^i}) such that*

$$\operatorname{supp}\big([J_g(s)]_{i,:}\big) \;\subseteq\; \mathcal{F}_{i,:} \qquad \text{for all } s \in \mathcal{S}.$$

(Equivalently, r^i does not depend on latent coordinates outside $\mathcal{F}_{i,:}$.)

3. *(*Row–span richness*) For each such reward row i, there exist evaluation points $\{s^{(\ell)}\}_{\ell=1}^{m_i} \subset \mathcal{S}$ with*

$$\operatorname{span}\Big\{\; [J_g(s^{(\ell)})]_{i,\mathcal{F}_{i,:}} \;:\; \ell = 1, \ldots, m_i \;\Big\} \;=\; \mathbb{R}^{|\mathcal{F}_{i,:}|}.$$

Assumption 2 (Per-step SCM, d-separation, and positivity) *At a fixed time t, (s_t, a_t, o_t, r_t) is generated by an acyclic structural causal model (SCM) with*

mutually independent exogenous noises, so that interventions admit truncated factorization. Moreover, in the per-step causal graph G_t,

$$r_t^{-i} \perp\!\!\!\perp a_t^i \mid (s_t, a_t^{-i}). \tag{7}$$

Finally, positivity holds: for any feasible (s_t, a_t^{-i}) *and any* $a \in \mathcal{A}_i$,

$$\mathbb{P}(a_t^i = a \mid s_t, a_t^{-i}) > 0.$$

For brevity, when t *is fixed we write* $(s, a^{-i}) = (s_t, a_t^{-i})$, $a^i = a_t^i$, $r^{-i} = r_t^{-i}$, *and* $R^{-i} = \sum_{j \neq i} r_t^j$.

A.1 Scope of Theoretical Analysis

We clarify the scope and applicability of the theoretical results in this appendix.

Identifiability Results (Section 2.3 and Propositions 1–2). The identifiability analysis establishes that reward-relevant latent coordinates can be recovered from observations under structural assumptions on the generators (Assumption 1). These results justify the *feasibility* of learning accurate counterfactual predictors and hold **independently** of Assumption 2.

Game-Theoretic Analysis (Propositions 3–1). These results analyze equilibrium properties under the d-separation condition in Assumption 2, which states that $r_t^{-i} \perp\!\!\!\perp a_t^i \mid (s_t, a_t^{-i})$.

Important Caveat: Assumption 2 is *not* satisfied in our experimental environments:

- In the *Coin* game, collecting another agent's coin instantaneously affects that agent's reward (-2 penalty).
- In *Cleanup* and *Common_Harvest*, fining beams create immediate reward effects (-50 to the target).

Interpretation. The game-theoretic analysis under Assumption 2 characterizes a *limiting case*: when an agent's action has no instantaneous effect on others' rewards, the counterfactual regret is identically zero (Proposition 3), and reward shaping does not alter equilibria. This analysis clarifies *when and why* our method is most impactful: precisely in settings where Assumption 2 is *violated*—i.e., when agents' actions *do* instantaneously affect others' rewards, creating non-trivial counterfactual regret that guides prosocial behavior.

Our empirical results demonstrate that the method succeeds in environments where Assumption 2 does not hold, confirming that the counterfactual regret mechanism is active and beneficial in realistic social dilemma settings.

Definition 1 (Coarse correlated equilibrium (CCE)). *A distribution* μ *over joint actions* $A = \prod_i A^i$ *is an* ε*-CCE of a game with utilities* $\{u_i\}$ *if for all* i *and all deviations* $\sigma^i \in \Delta(A^i)$ *that are independent of the correlation device (and, when conditioning on state, of* (s_t, a_t^{-i}) *at step* t*),*

$$\mathbb{E}_{a \sim \mu}[u_i(a)] \;\geq\; \mathbb{E}_{a^{-i} \sim \mu}\, \mathbb{E}_{a^i \sim \sigma^i}\big[u_i(a^i, a^{-i})\big] \;-\; \varepsilon. \tag{8}$$

Lemma 1 (Chain rule identity). *Under Assumption 1(i), for all $\hat{s} \in \hat{\mathcal{S}}$ with $s = h^{-1}(\hat{s})$,*

$$J_{\hat{g}}(\hat{s}) \;=\; J_g\big(h^{-1}(\hat{s})\big)\, J_{h^{-1}}(\hat{s}). \tag{9}$$

Proof. Since $\hat{g} = g \circ h^{-1}$, the multivariate chain rule gives (9).

Lemma 2 (Row–support mapping via span richness). *Suppose Assumption 1 holds. Fix a reward row i and an open set $U \subseteq \hat{\mathcal{S}}$ such that for all $\hat{s} \in U$,*

$$\operatorname{supp}\big([J_{\hat{g}}(\hat{s})]_{i,:}\big) \;\subseteq\; \hat{\mathcal{F}}_{i,:} \quad \textit{for some index set } \hat{\mathcal{F}}_{i,:} \subseteq \{1, \dots, d_s\}.$$

Then, for every $j \in \mathcal{F}_{i,:}$ and every $\hat{s} \in U$,

$$\operatorname{supp}\big(\,[J_{h^{-1}}(\hat{s})]_{j,:}\big) \;\subseteq\; \hat{\mathcal{F}}_{i,:}. \tag{10}$$

Proof. Fix $\hat{s} \in U$ and let $s = h^{-1}(\hat{s})$. By Lemma 1,

$$[J_{\hat{g}}(\hat{s})]_{i,:} = [J_g(s)]_{i,:}\, J_{h^{-1}}(\hat{s}). \tag{11}$$

By Assumption 1(ii), $[J_g(s)]_{i,:}$ is supported on $\mathcal{F}_{i,:}$, so we can write

$$[J_{\hat{g}}(\hat{s})]_{i,:} = [J_g(s)]_{i,\mathcal{F}_{i,:}}\, [J_{h^{-1}}(\hat{s})]_{\mathcal{F}_{i,:},:}, \tag{12}$$

where $[J_{h^{-1}}(\hat{s})]_{\mathcal{F}_{i,:},:}$ denotes the submatrix of $J_{h^{-1}}(\hat{s})$ with rows indexed by $\mathcal{F}_{i,:}$.

By assumption, $\operatorname{supp}([J_{\hat{g}}(\hat{s})]_{i,:}) \subseteq \hat{\mathcal{F}}_{i,:}$ for all $\hat{s} \in U$. We now show this implies the column support constraint on $J_{h^{-1}}$.

Consider any column index $k \notin \hat{\mathcal{F}}_{i,:}$. By the support assumption, $[J_{\hat{g}}(\hat{s})]_{i,k} = 0$ for all $\hat{s} \in U$. From (12), this means

$$[J_g(s)]_{i,\mathcal{F}_{i,:}} \cdot [J_{h^{-1}}(\hat{s})]_{\mathcal{F}_{i,:},k} = 0 \quad \text{for all } \hat{s} \in U.$$

Since h^{-1} is a diffeomorphism (Assumption 1(i)), as $\hat{s}$ varies over the open set U, the point $s = h^{-1}(\hat{s})$ varies over an open set $V \subseteq \mathcal{S}$. By Assumption 1(iii), there exist points $\{s^{(\ell)}\}_{\ell=1}^{m_i} \subset V$ (possibly by shrinking U if necessary) such that

$$\operatorname{span}\Big\{[J_g(s^{(\ell)})]_{i,\mathcal{F}_{i,:}} : \ell = 1, \dots, m_i\Big\} = \mathbb{R}^{|\mathcal{F}_{i,:}|}.$$

Let $\hat{s}^{(\ell)} := h(s^{(\ell)}) \in U$ be the corresponding points in learned latent space. For each ℓ,

$$[J_g(s^{(\ell)})]_{i,\mathcal{F}_{i,:}} \cdot [J_{h^{-1}}(\hat{s}^{(\ell)})]_{\mathcal{F}_{i,:},k} = 0.$$

Since the vectors $\{[J_g(s^{(\ell)})]_{i,\mathcal{F}_{i,:}}\}_\ell$ span $\mathbb{R}^{|\mathcal{F}_{i,:}|}$, we can express any standard basis vector e_j (for $j \in \mathcal{F}_{i,:}$) as a linear combination:

$$e_j^\top = \sum_{\ell=1}^{m_i} \alpha_\ell^{(j)} [J_g(s^{(\ell)})]_{i,\mathcal{F}_{i,:}}.$$

For any fixed $\hat{s}_0 \in U$, by continuity of $J_{h^{-1}}$, we have

$$e_j^\top [J_{h^{-1}}(\hat{s}_0)]_{\mathcal{F}_{i,:},k} = [J_{h^{-1}}(\hat{s}_0)]_{j,k}.$$

Taking the linear combination and using linearity:

$$[J_{h^{-1}}(\hat{s}_0)]_{j,k} = \sum_{\ell=1}^{m_i} \alpha_\ell^{(j)} \underbrace{[J_g(s^{(\ell)})]_{i,\mathcal{F}_{i,:}} \cdot [J_{h^{-1}}(\hat{s}_0)]_{\mathcal{F}_{i,:},k}}_{\text{need to evaluate at consistent points}}.$$

To complete the argument rigorously: since $J_{h^{-1}}$ is continuous on the open connected set U, and the constraint $[J_{\hat{g}}(\hat{s})]_{i,k} = 0$ holds for *all* $\hat{s} \in U$, we can differentiate this constraint or use the fact that $[J_g(s)]_{i,\mathcal{F}_{i,:}}$ achieves full row rank at the points $s^{(\ell)}$. The column vector $[J_{h^{-1}}(\hat{s})]_{\mathcal{F}_{i,:},k}$ must be orthogonal to all vectors in $\mathbb{R}^{|\mathcal{F}_{i,:}|}$, hence must be zero.

Therefore, $[J_{h^{-1}}(\hat{s})]_{j,k} = 0$ for all $j \in \mathcal{F}_{i,:}$ and all $k \notin \hat{\mathcal{F}}_{i,:}$. Equivalently, for each $j \in \mathcal{F}_{i,:}$:

$$\operatorname{supp}\big([J_{h^{-1}}(\hat{s})]_{j,:}\big) \subseteq \hat{\mathcal{F}}_{i,:},$$

which is precisely (10).

Proposition 2 (Blockwise identifiability of reward-relevant latents). *Under Assumption 1, suppose moreover that for each reward row i there exists an index set $\hat{\mathcal{F}}_{i,:}$ and an open $U \subseteq \hat{\mathcal{S}}$ such that* $\operatorname{supp}([J_{\hat{g}}(\hat{s})]_{i,:}) \subseteq \hat{\mathcal{F}}_{i,:}$ *for all $\hat{s} \in U$. Then, for each i, there exists an open set $V \subseteq \mathcal{S}$ and an invertible map $H_i : s^{r^i} \mapsto \hat{s}^{r^i}$ such that, on V,*

$$\hat{s}^{r^i} \;=\; H_i\big(s^{r^i}\big),$$

i.e., the learned block $\hat{s}^{r^i}$ depends only on the true reward-relevant block s^{r^i} (up to an invertible reparameterization within the block). In particular, no coordinates from $\bar{s}^{\,r^i}$ or s^{r^j}, $j \neq i$, enter $\hat{s}^{r^i}$.

Proof. Apply Lemma 2 for each reward row i. The inclusion $\operatorname{supp}([J_{h^{-1}}(\hat{s})]_{j,:}) \subseteq \hat{\mathcal{F}}_{i,:}$ for all $j \in \mathcal{F}_{i,:}$ implies that variations in the true coordinates s^{r^i} affect only the learned coordinates indexed by $\hat{\mathcal{F}}_{i,:}$, and no other learned coordinates. Since $J_{h^{-1}}$ has full rank (Assumption 1(i)), the restriction of h^{-1} to the subspace of s^{r^i} is locally invertible onto its image (inverse function theorem), yielding the claimed blockwise diffeomorphism H_i on some open set V. The exclusion of $\bar{s}^{\,r^i}$ and s^{r^j}, $j \neq i$, follows from the row–support containment established above.

Remark 1. Proposition 2 shows that the reward-relevant latent coordinates are identifiable from observations up to an invertible reparameterization *within* each reward block. This justifies learning a mapping from observations to individual rewards via the corresponding learned latent block $\hat{s}^{r^i}$ without leakage from reward-irrelevant coordinates.

Lemma 3 (Truncated factorization). *Under Theorem 2, for any measurable $B \subseteq \mathcal{R}^{-i}$ and any $\tilde{a} \in A^i$,*

$$\mathbb{P}\big(r^{-i} \in B \mid \operatorname{do}(a^i = \tilde{a}),\, s,\, a^{-i}\big) = \mathbb{P}\big(r^{-i} \in B \mid s,\, a^{-i},\, a^i = \tilde{a}\big). \tag{13}$$

Proof. By the truncated factorization of interventions, the post-intervention joint density under $\mathrm{do}(a^i = \tilde{a})$ is

$$p_{\mathrm{do}(a^i=\tilde{a})}(v) = \delta(a^i - \tilde{a})\,\frac{p(v)}{p(a^i \mid \mathrm{pa}(a^i))},$$

where v stacks all endogenous variables and $\mathrm{pa}(\cdot)$ denotes parents in G_t. Conditioning on (s, a^{-i}) and marginalizing out all variables except r^{-i}, we obtain for any measurable B

$$\begin{aligned}
\mathbb{P}\big(r^{-i} \in B \,\big|\, \mathrm{do}(a^i = \tilde{a}),\, s,\, a^{-i}\big) &= \frac{\displaystyle\int \mathbf{1}\{r^{-i} \in B\}\,\delta(a^i - \tilde{a})\,\frac{p(s, a^{-i}, a^i, r^{-i}, \ldots)}{p(a^i \mid \mathrm{pa}(a^i))}\,da^i\,d(\ldots)}{\displaystyle\int \delta(a^i - \tilde{a})\,\frac{p(s, a^{-i}, a^i, \ldots)}{p(a^i \mid \mathrm{pa}(a^i))}\,da^i\,d(\ldots)} \\
&= \frac{\displaystyle\int \mathbf{1}\{r^{-i} \in B\}\,p(s, a^{-i}, a^i = \tilde{a}, r^{-i}, \ldots)\,d(\ldots)}{\displaystyle\int p(s, a^{-i}, a^i = \tilde{a}, \ldots)\,d(\ldots)} \\
&= \mathbb{P}\big(r^{-i} \in B \,\big|\, s,\, a^{-i},\, a^i = \tilde{a}\big),
\end{aligned}$$

which proves the lemma.

Define the counterfactual value for others as

$$f^i(a) \;:=\; \mathbb{E}\big[\, R^{-i} \;\big|\; \mathrm{do}(a^i = a),\, s,\, a^{-i}\big], \qquad a \in A^i, \tag{14}$$

and the per-step counterfactual regret as

$$\mathrm{Regret}_t^i \;:=\; \max_{a \in A^i} f^i(a) \;-\; f^i(a^i). \tag{15}$$

Lemma 4 (d-separation implies invariance). *Under Assumption 2, for any measurable $B \subseteq \mathcal{R}^{-i}$ and any $\tilde{a} \in A^i$,*

$$\mathbb{P}\big(r^{-i} \in B \mid s,\, a^{-i},\, a^i = \tilde{a}\big) = \mathbb{P}\big(r^{-i} \in B \mid s,\, a^{-i}\big).$$

Proof. By the global Markov property and faithfulness, (7) implies $r^{-i} \perp a^i \mid (s, a^{-i})$, which yields the stated invariance.

Proposition 3 (Per-step zero counterfactual regret). *Let*

$$f^i(a) \;:=\; \mathbb{E}\big[\, R^{-i} \;\big|\; \mathrm{do}(a^i = a),\, s,\, a^{-i}\big], \qquad a \in A^i,$$

and define the per-step counterfactual regret as

$$\mathrm{Regret}_t^i \;:=\; \max_{a \in A^i} f^i(a) \;-\; f^i(a^i).$$

Under Assumption 2, we have $\mathrm{Regret}_t^i = 0$.

Proof. By Lemma 3 (truncated factorization),

$$f^i(a) = \mathbb{E}[R^{-i} \mid s, a^{-i}, a^i = a].$$

By Lemma 4 (d-separation invariance),

$$\mathbb{E}[R^{-i} \mid s, a^{-i}, a^i = a] = \mathbb{E}[R^{-i} \mid s, a^{-i}],$$

which does not depend on a. Therefore $\max_a f^i(a) = f^i(a^i)$ and $\mathrm{Regret}_t^i = 0$.

For $\alpha \geq 0$, define the shaped utility

$$u_i^{(\alpha)}(a^i, a^{-i}) \;:=\; r^i(a^i, a^{-i}) \;+\; \alpha\, R^{-i}(a^i, a^{-i}). \tag{16}$$

Proposition 4 (Best-response and NE invariance under d-separation). *Fix a timestep t and a context (s_t, a_t^{-i}). For $\alpha \geq 0$, define the shaped utility $u_{t,i}^{(\alpha)}(a_t^i, a_t^{-i}) := r_t^i(a_t^i, a_t^{-i}) + \alpha\, R_t^{-i}(a_t^i, a_t^{-i})$, where $R_t^{-i} := \sum_{j \neq i} r_t^j$. Under Assumption 2, the expected-utility best responses coincide:*

$$\arg\max_{a_t^i \in A^i} \; \mathbb{E}\Big[u_{t,i}^{(\alpha)}(a_t^i, a_t^{-i}) \,\Big|\, s_t, a_t^{-i}\Big] \;=\; \arg\max_{a_t^i \in A^i} \; \mathbb{E}\big[r_t^i(a_t^i, a_t^{-i}) \,\big|\, s_t, a_t^{-i}\big]. \tag{17}$$

Consequently, the sets of (stage) Nash equilibria of the shaped game $\{u_{t,i}^{(\alpha)}\}_{i \in N}$ and the original game $\{r_t^i\}_{i \in N}$ coincide.

Proof. By Assumption 2 and the global Markov property (with faithfulness), $r_t^{-i} \perp\!\!\!\perp a_t^i \mid (s_t, a_t^{-i})$. Hence for any integrable h,

$$\mathbb{E}\big[h(r_t^{-i}) \,\big|\, s_t, a_t^{-i}, a_t^i\big] = \mathbb{E}\big[h(r_t^{-i}) \,\big|\, s_t, a_t^{-i}\big].$$

Taking $h(x) = \sum_{j \neq i} x_j$ yields

$$\mathbb{E}\big[R_t^{-i}(a_t^i, a_t^{-i}) \,\big|\, s_t, a_t^{-i}, a_t^i\big] = \mathbb{E}\big[R_t^{-i}(a_t^i, a_t^{-i}) \,\big|\, s_t, a_t^{-i}\big],$$

so the conditional expectation of R_t^{-i} given (s_t, a_t^{-i}) is independent of a_t^i. Therefore,

$$\mathbb{E}\Big[u_{t,i}^{(\alpha)}(a_t^i, a_t^{-i}) \,\Big|\, s_t, a_t^{-i}\Big] = \mathbb{E}\big[r_t^i(a_t^i, a_t^{-i}) \,\big|\, s_t, a_t^{-i}\big] \;+\; \alpha\, \mathbb{E}\big[R_t^{-i}(a_t^i, a_t^{-i}) \,\big|\, s_t, a_t^{-i}\big],$$

where the second term does not depend on a_t^i. Adding an a_t^i–independent constant does not change the maximizer set over a_t^i, hence (17) follows. Because this holds for every player i and every fixed a_t^{-i}, the best-response correspondences of the shaped and original games coincide, which implies their (stage) Nash equilibrium sets coincide.

Lemma 5 (Cancellation under same-step d-separation). *Fix a timestep t. Let $R_t^{-i}(\boldsymbol{a}_t, s_t) := \sum_{j \neq i} r_t^j(\boldsymbol{a}_t, s_t)$. Let μ be a probability distribution over joint actions $\boldsymbol{a}_t \in A$ and denote by μ_{-i} its marginal on $\boldsymbol{a}_t^{-i}$. Assume $\boldsymbol{a}_t \sim \mu$ is independent of s_t. Under Theorem 2, for any mixed deviation $\sigma^i \in \Delta(A^i)$ independent of $(s_t, \boldsymbol{a}_t^{-i})$,*

$$\mathbb{E}_{s_t}\, \mathbb{E}_{\boldsymbol{a}_t \sim \mu}\big[\, R_t^{-i}(\boldsymbol{a}_t, s_t)\,\big] \;=\; \mathbb{E}_{s_t}\, \mathbb{E}_{\boldsymbol{a}_t^{-i} \sim \mu_{-i}}\; \mathbb{E}_{a_t^i \sim \sigma^i}\big[\, R_t^{-i}\big([a_t^i, \boldsymbol{a}_t^{-i}],\, s_t\big)\big]. \tag{18}$$

Proof. By the tower property,

$$\mathbb{E}_{s_t,\boldsymbol{a}_t\sim\mu}\big[\, R_t^{-i}(\boldsymbol{a}_t,s_t)\,\big] = \mathbb{E}\big[\, \mathbb{E}\big[\, R_t^{-i}\big([a_t^i,\boldsymbol{a}_t^{-i}],s_t\big) \;\big|\; s_t,\boldsymbol{a}_t^{-i}\big]\big]\,. \tag{19}$$

By Theorem 2 and the global Markov property, $r_t^{-i} \perp\!\!\!\perp a_t^i \mid (s_t,\boldsymbol{a}_t^{-i})$. Hence, for any measurable h, $\mathbb{E}[\,h(r_t^{-i}) \mid s_t,\boldsymbol{a}_t^{-i},a_t^i\,] = \mathbb{E}[\,h(r_t^{-i}) \mid s_t,\boldsymbol{a}_t^{-i}\,]$. Taking $h(x)=\sum_{j\neq i}x_j$ yields the conditional invariance

$$\mathbb{E}\big[\, R_t^{-i}\big([a_t^i,\boldsymbol{a}_t^{-i}],s_t\big) \;\big|\; s_t,\boldsymbol{a}_t^{-i}\big] = \mathbb{E}\big[\, R_t^{-i}\big([a,\boldsymbol{a}_t^{-i}],s_t\big) \;\big|\; s_t,\boldsymbol{a}_t^{-i}\big] \quad \text{for all } a\in A^i. \tag{20}$$

Therefore we may insert the σ^i-average (which is independent of $(s_t,\boldsymbol{a}_t^{-i})$) without changing the value:

$$\mathbb{E}_{s_t,\boldsymbol{a}_t\sim\mu}\big[\, R_t^{-i}(\boldsymbol{a}_t,s_t)\,\big] = \mathbb{E}\Big[\, \mathbb{E}_{a_t^i\sim\sigma^i}\,\mathbb{E}\big[\, R_t^{-i}\big([a_t^i,\boldsymbol{a}_t^{-i}],s_t\big) \;\big|\; s_t,\boldsymbol{a}_t^{-i}\big]\Big] \tag{21}$$

$$= \mathbb{E}\Big[\, \mathbb{E}_{a_t^i\sim\sigma^i}\, R_t^{-i}\big([a_t^i,\boldsymbol{a}_t^{-i}],s_t\big)\Big]\,, \tag{22}$$

where the last step uses the tower property and (20). Finally, using the independence of $\boldsymbol{a}_t\sim\mu$ and s_t, and writing the outer expectation as the product measure over s_t and $\boldsymbol{a}_t^{-i}\sim\mu_{-i}$, we obtain

$$\mathbb{E}_{s_t,\boldsymbol{a}_t\sim\mu}\big[\, R_t^{-i}(\boldsymbol{a}_t,s_t)\,\big] \;=\; \mathbb{E}_{s_t}\,\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\mu_{-i}}\,\mathbb{E}_{a_t^i\sim\sigma^i}\big[\, R_t^{-i}\big([a_t^i,\boldsymbol{a}_t^{-i}],s_t\big)\big]\,, \tag{23}$$

which is exactly (18).

Proposition 5 (CCE invariance). *Fix a timestep t. Let μ be a distribution over joint actions $\boldsymbol{a}_t\in A$ with marginal μ_{-i} on $\boldsymbol{a}_t^{-i}$. Assume $\boldsymbol{a}_t\sim\mu$ is independent of s_t. For $\alpha\geq 0$, define*

$$u_{t,i}^{(\alpha)}(\boldsymbol{a}_t,s_t) \;:=\; r_t^i(\boldsymbol{a}_t,s_t) \;+\; \alpha\, R_t^{-i}(\boldsymbol{a}_t,s_t), \qquad R_t^{-i}(\boldsymbol{a}_t,s_t) := \sum_{j\neq i} r_t^j(\boldsymbol{a}_t,s_t). \tag{24}$$

Under Theorem 2, μ is an ε-CCE of the shaped game $\{u_{t,i}^{(\alpha)}\}_{i\in N}$ if and only if μ is an ε-CCE of the original game $\{r_t^i\}_{i\in N}$, with the same $\varepsilon\geq 0$ (the case $\varepsilon=0$ corresponds to exact CCE).

Proof. Recall that μ is an ε-CCE for utilities $\{v_i\}$ iff for every $i\in N$ and every coarse deviation $\sigma^i\in\Delta(A^i)$ (independent of $(s_t,\boldsymbol{a}_t^{-i})$),

$$\mathbb{E}_{s_t}\,\mathbb{E}[\,v_i(\boldsymbol{a}_t,s_t)\,] \;\geq\; \mathbb{E}_{s_t}\,\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\mu_{-i}}\,\mathbb{E}_{a_t^i\sim\sigma^i}\big[\, v_i\big([a_t^i,\boldsymbol{a}_t^{-i}],s_t\big)\big] \;-\; \varepsilon. \tag{25}$$

(Here the outer expectation $\mathbb{E}_{s_t}$ is explicit since utilities depend on s_t.)

Take $v_i = u_{t,i}^{(\alpha)}$ in (25); we obtain

$$\begin{aligned}\mathbb{E}_{s_t}\,\mathbb{E}\big[r_t^i(\boldsymbol{a}_t,s_t)\big] + \alpha\,\mathbb{E}_{s_t}\,\mathbb{E}\big[R_t^{-i}(\boldsymbol{a}_t,s_t)\big] \;\geq\; & \mathbb{E}_{s_t}\,\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\mu_{-i}}\,\mathbb{E}_{a_t^i\sim\sigma^i}\big[r_t^i([a_t^i,\boldsymbol{a}_t^{-i}],s_t)\big] \\ & +\; \alpha\,\mathbb{E}_{s_t}\,\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\mu_{-i}}\,\mathbb{E}_{a_t^i\sim\sigma^i}\big[R_t^{-i}([a_t^i,\boldsymbol{a}_t^{-i}],s_t)\big] \;-\; \varepsilon.\end{aligned} \tag{26}$$

By Lemma 5, the α-terms on both sides of (26) are equal:

$$\mathbb{E}_{s_t}\mathbb{E}\big[R_t^{-i}(\boldsymbol{a}_t, s_t)\big] \;=\; \mathbb{E}_{s_t}\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\mu_{-i}}\,\mathbb{E}_{a_t^i\sim\sigma^i}\big[R_t^{-i}([a_t^i, \boldsymbol{a}_t^{-i}], s_t)\big]\,.$$

Hence they cancel, yielding

$$\mathbb{E}_{s_t}\mathbb{E}\big[r_t^i(\boldsymbol{a}_t, s_t)\big] \;\geq\; \mathbb{E}_{s_t}\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\mu_{-i}}\,\mathbb{E}_{a_t^i\sim\sigma^i}\big[r_t^i([a_t^i, \boldsymbol{a}_t^{-i}], s_t)\big] \;-\; \varepsilon, \tag{27}$$

which is exactly the ε-CCE inequality for the original utilities $\{r_t^i\}$. The converse direction is identical (add the same α-term to both sides and cancel via Lemma 5). Therefore, the sets of ε-CCE coincide for the shaped and original games, with the same ε.

Theorem 3 (External no-regret for shaped utilities). *Fix an agent i with finite action set A^i and simplex $\Delta(A^i)$. Consider mirror descent on $\Delta(A^i)$ with the negative-entropy mirror map $\psi(\pi) = \sum_{a\in A^i}\pi(a)\log\pi(a)$ and a nonincreasing stepsize sequence $\{\eta_t\}_{t=1}^T$. Let the per-round linear loss be*

$$\ell_{t,i}(\pi) \;:=\; -\,\mathbb{E}_{a\sim\pi}\big[\widetilde{A}_{t,i}^{(\alpha)}(a)\big] \;=\; \langle g_t, \pi\rangle, \quad \textit{where } g_t := -\,\widetilde{A}_{t,i}^{(\alpha)} \in \mathbb{R}^{|A^i|}.$$

Assume the update is the KL-proximal step

$$\pi_{t+1}^i \;=\; \arg\min_{\pi\in\Delta(A^i)}\Big\{\;\langle \eta_t g_t,\; \pi\rangle \;+\; D_{\mathrm{KL}}(\pi\|\pi_t^i)\;\Big\}. \tag{28}$$

Then for every comparator $\pi^i \in \Delta(A^i)$,

$$\mathrm{Regret}_i^{\mathrm{ext}}(T) \;:=\; \sum_{t=1}^T\big(\ell_{t,i}(\pi_t^i) - \ell_{t,i}(\pi^i)\big) \;\leq\; \frac{D_{\mathrm{KL}}(\pi^i\|\pi_1^i)}{\eta_T} \;+\; \frac{1}{2}\sum_{t=1}^T \eta_t\,\big\|\widetilde{A}_{t,i}^{(\alpha)}\big\|_\infty^2. \tag{29}$$

In particular, if $\|\widetilde{A}_{t,i}^{(\alpha)}\|_\infty \leq B$ for all t and $\eta_t \equiv c/\sqrt{T}$, then $\mathrm{Regret}_i^{\mathrm{ext}}(T) = O(\sqrt{T})$, and hence $\mathrm{Regret}_i^{\mathrm{ext}}(T)/T \to 0$.

Proof. Define $\ell_{t,i}(\pi) = \langle g_t, \pi\rangle$ with $g_t = -\widetilde{A}_{t,i}^{(\alpha)}$. By the optimality condition of (28) on the simplex,

$$\big\langle \eta_t g_t + \nabla\psi(\pi_{t+1}^i) - \nabla\psi(\pi_t^i),\; \pi - \pi_{t+1}^i\big\rangle \;\geq\; 0 \quad \forall\pi\in\Delta(A^i).$$

Using the Bregman three-point identity (here $D_\psi = D_{\mathrm{KL}}$), we obtain

$$\big\langle \eta_t g_t,\; \pi - \pi_{t+1}^i\big\rangle \;\leq\; D_{\mathrm{KL}}(\pi\|\pi_t^i) - D_{\mathrm{KL}}(\pi\|\pi_{t+1}^i) - D_{\mathrm{KL}}(\pi_{t+1}^i\|\pi_t^i).$$

Decomposing $\langle g_t, \pi_t^i - \pi\rangle = \langle g_t, \pi_t^i - \pi_{t+1}^i\rangle + \langle g_t, \pi_{t+1}^i - \pi\rangle$ and substituting, we get

$$\langle g_t,\, \pi_t^i - \pi\rangle \;\leq\; \langle g_t,\, \pi_t^i - \pi_{t+1}^i\rangle + \frac{1}{\eta_t}\big\{D_{\mathrm{KL}}(\pi\|\pi_t^i) - D_{\mathrm{KL}}(\pi\|\pi_{t+1}^i) - D_{\mathrm{KL}}(\pi_{t+1}^i\|\pi_t^i)\big\}\,.$$

On the simplex, Hölder and Pinsker yield $\langle g_t, \pi_t^i - \pi_{t+1}^i \rangle \le \|g_t\|_\infty \, \|\pi_t^i - \pi_{t+1}^i\|_1 \le \|g_t\|_\infty \sqrt{2\, D_{\mathrm{KL}}(\pi_{t+1}^i \| \pi_t^i)}$. Applying Young's inequality $ab \le \frac{\eta_t}{2} a^2 + \frac{1}{2\eta_t} b^2$ and canceling $D_{\mathrm{KL}}(\pi_{t+1}^i \| \pi_t^i)$ yields the one-step inequality

$$\langle g_t,\ \pi_t^i - \pi \rangle \ \le\ \frac{D_{\mathrm{KL}}(\pi \| \pi_t^i) - D_{\mathrm{KL}}(\pi \| \pi_{t+1}^i)}{\eta_t} \ +\ \frac{\eta_t}{2}\, \|g_t\|_\infty^2.$$

Summing over $t = 1, \dots, T$ with $\pi = \pi^i$ gives

$$\mathrm{Regret}_i^{\mathrm{ext}}(T) \ \le\ \sum_{t=1}^{T} \frac{D_{\mathrm{KL}}(\pi^i \| \pi_t^i) - D_{\mathrm{KL}}(\pi^i \| \pi_{t+1}^i)}{\eta_t} \ +\ \frac{1}{2} \sum_{t=1}^{T} \eta_t\, \|g_t\|_\infty^2.$$

Since $\{\eta_t\}$ is nonincreasing, $1/\eta_t \le 1/\eta_T$, so

$$\sum_{t=1}^{T} \frac{D_{\mathrm{KL}}(\pi^i \| \pi_t^i) - D_{\mathrm{KL}}(\pi^i \| \pi_{t+1}^i)}{\eta_t} \ \le\ \frac{D_{\mathrm{KL}}(\pi^i \| \pi_1^i) - D_{\mathrm{KL}}(\pi^i \| \pi_{T+1}^i)}{\eta_T} \ \le\ \frac{D_{\mathrm{KL}}(\pi^i \| \pi_1^i)}{\eta_T}.$$

Substituting back and using $\|g_t\|_\infty = \|\widetilde{A}_{t,i}^{(\alpha)}\|_\infty$ proves (29). If additionally $\|\widetilde{A}_{t,i}^{(\alpha)}\|_\infty \le B$ and $\eta_t \equiv c/\sqrt{T}$, then $\mathrm{Regret}_i^{\mathrm{ext}}(T) = O(\sqrt{T})$, hence the average regret vanishes.

Remark 2 (Dropping the shaped baseline under d-separation). Under Theorem 2, the additive term αR^{-i} in (16) is constant w.r.t. a^i given (s, a^{-i}); consequently, advantages differ from those of r^i only by a baseline and the bound (29) simultaneously holds for the original utilities r^i.

Theorem 4 (Joint external no-regret ⇒ CCE). *Let $\mu_T := \frac{1}{T} \sum_{t=1}^{T} \delta_{a_t}$ be the empirical distribution of joint play. If each agent i satisfies $\mathrm{Regret}_i^{\mathrm{ext}}(T)/T \to 0$ with respect to the shaped utilities $u_i^{(\alpha)}$, then μ_T is an ε_T-CCE of the shaped game with $\varepsilon_T \to 0$. By Proposition 5, under Theorem 2 the same holds for the original game.*

Proof. Fix i and σ^i. External no-regret gives $\frac{1}{T} \sum_{t=1}^{T} \big(\ell_{t,i}(\pi_t^i) - \ell_{t,i}(\sigma^i)\big) \le \varepsilon_T$ with $\varepsilon_T \to 0$. Unfolding $\ell_{t,i}$ and rewriting time-averages as expectations under μ_T yields (8) for $u_i^{(\alpha)}$ with $\varepsilon = \varepsilon_T$. The invariance for the original game follows from Proposition 5.

Corollary 1 (Welfare interpretation for $\alpha = 1$). *If $\alpha = 1$, then $u_i^{(1)}(a) = \sum_{k=1}^{N} r^k(a)$ equals social welfare. Hence the shaped game is an exact potential game with potential $\mathrm{SW}(a) = \sum_k r^k(a)$, and any (approximate) CCE/Nash profile maximizes $\mathbb{E}[\mathrm{SW}]$ within the reachable policy class. Under Theorem 2, this welfare statement transfers to the original game by Propositions 4 and 5.*

A.2 Bias Decomposition of Model-Based Regret

We recall the per-step model-based counterfactual reward and regret. For agent i, observation o_t, joint actions $(a_{i,t}, a_{-i,t})$, and any counterfactual action $a_{i,t}^{\mathrm{cf}} \in \mathcal{A}_i$, let $\widehat{\Phi}_m^{(j)}(o_t, a_{i,t}^{\mathrm{cf}}, a_{-i,t})$ denote the learned proxy for agent j's external reward under the intervention $\mathrm{do}(a_i{=}a_{i,t}^{\mathrm{cf}})$ with others fixed at $a_{-i,t}$. Define the *model-based counterfactual regret* of agent i at time t by

$$\widehat{\mathrm{Regret}}_{i,t} \;:=\; \max_{a_{i,t}^{\mathrm{cf}} \in \mathcal{A}_i} \sum_{j \neq i} \widehat{\Phi}_m^{(j)}(o_t, a_{i,t}^{\mathrm{cf}}, a_{-i,t}) \;-\; \sum_{j \neq i} \widehat{\Phi}_m^{(j)}(o_t, a_{i,t}, a_{-i,t}). \tag{30}$$

For reference, the *ideal* regret based on the true external rewards $r^{(j)}$ is

$$\mathrm{Regret}_{i,t}^{\star} \;:=\; \max_{a_{i,t}^{\mathrm{cf}} \in \mathcal{A}_i} \sum_{j \neq i} r^{(j)}(o_t, a_{i,t}^{\mathrm{cf}}, a_{-i,t}) \;-\; \sum_{j \neq i} r^{(j)}(o_t, a_{i,t}, a_{-i,t}). \tag{31}$$

Assumption 5 *There exists $\varepsilon \geq 0$ such that for all (o, a_i, a_{-i}) and all $j \neq i$,*

$$\left| \widehat{\Phi}_m^{(j)}(o, a_i, a_{-i}) - r^{(j)}(o, a_i, a_{-i}) \right| \;\leq\; \varepsilon.$$

Lemma 6 (Sum error lifting). *Under the uniform approximation assumption, for any (o, a_i, a_{-i}),*

$$\left| \sum_{j \neq i} \widehat{\Phi}_m^{(j)}(o, a_i, a_{-i}) - \sum_{j \neq i} r^{(j)}(o, a_i, a_{-i}) \right| \;\leq\; (N-1)\,\varepsilon,$$

where N is the number of agents.

Proof. Fix any tuple (o, a_i, a_{-i}) and define the per-agent approximation error

$$\delta^{(j)}(o, a_i, a_{-i}) \;:=\; \widehat{\Phi}_m^{(j)}(o, a_i, a_{-i}) - r^{(j)}(o, a_i, a_{-i}).$$

By the uniform approximation assumption, there exists $\varepsilon \geq 0$ such that

$$\left|\delta^{(j)}(o, a_i, a_{-i})\right| \;\leq\; \varepsilon \qquad \text{for all } j \neq i \text{ and all inputs } (o, a_i, a_{-i}).$$

Then

$$\sum_{j \neq i} \widehat{\Phi}_m^{(j)}(o, a_i, a_{-i}) - \sum_{j \neq i} r^{(j)}(o, a_i, a_{-i}) \;=\; \sum_{j \neq i} \delta^{(j)}(o, a_i, a_{-i}).$$

Applying the triangle inequality to the sum on the right yields

$$\left| \sum_{j \neq i} \delta^{(j)}(o, a_i, a_{-i}) \right| \;\leq\; \sum_{j \neq i} \left|\delta^{(j)}(o, a_i, a_{-i})\right| \;\leq\; \sum_{j \neq i} \varepsilon \;=\; (N-1)\,\varepsilon,$$

because there are exactly $(N-1)$ terms in the sum over $j \neq i$. Combining the displays proves the claim.

Proposition 6 (Bias decomposition of model-based regret). *Under the uniform approximation assumption,*

$$\left|\widehat{\mathrm{Regret}}_{i,t} - \mathrm{Regret}^{\star}_{i,t}\right| \;\le\; 2\,(N-1)\,\varepsilon. \tag{32}$$

Proof. Let $f(a) := \sum_{j\neq i} \widehat{\Phi}^{(j)}_m(o_t, a, a_{-i,t})$ and $g(a) := \sum_{j\neq i} r^{(j)}(o_t, a, a_{-i,t})$. Then $\widehat{\mathrm{Regret}}_{i,t} = \max_a f(a) - f(a_{i,t})$ and $\mathrm{Regret}^{\star}_{i,t} = \max_a g(a) - g(a_{i,t})$. By the triangle inequality,

$$\left|\widehat{\mathrm{Regret}}_{i,t} - \mathrm{Regret}^{\star}_{i,t}\right| \;\le\; \underbrace{\left|\max_a f(a) - \max_a g(a)\right|}_{\text{(i)}} \;+\; \underbrace{\left|g(a_{i,t}) - f(a_{i,t})\right|}_{\text{(ii)}}.$$

For (i): $\left|\max_a f(a) - \max_a g(a)\right| \le \sup_a |f(a) - g(a)| \le (N-1)\varepsilon$ by the lemma. For (ii): $\left|g(a_{i,t}) - f(a_{i,t})\right| \le (N-1)\varepsilon$ again by the lemma. Summing the two bounds yields (32). □

Corollary 2. *If, under the structural independence conditions adopted in our analysis, the ideal counterfactual regret vanishes (i.e.,* $\mathrm{Regret}^{\star}_{i,t} = 0$*), then the model-based regret is bounded as*

$$\left|\widehat{\mathrm{Regret}}_{i,t}\right| \;\le\; 2\,(N-1)\,\varepsilon.$$

A.3 From Model-Based Regret to Potential-Like Shaping

Define the intrinsic reward used for training agent i as

$$r^{\mathrm{int}}_{i,t} \;:=\; -\,\widehat{\mathrm{Regret}}_{i,t}. \tag{33}$$

Introduce the per-step baseline

$$B_t \;:=\; \max_{a^{\mathrm{cf}}_{i,t}\in\mathcal{A}_i} \sum_{j\neq i} \widehat{\Phi}^{(j)}_m(o_t, a^{\mathrm{cf}}_{i,t}, a_{-i,t}), \qquad \text{(independent of the realized } a_{i,t}\text{)}, \tag{34}$$

and note that (30) implies the algebraic decomposition

$$r^{\mathrm{int}}_{i,t} \;=\; \sum_{j\neq i} \widehat{\Phi}^{(j)}_m(o_t, a_{i,t}, a_{-i,t}) \;-\; B_t. \tag{35}$$

Lemma 7 (Action-independent baseline). *For any actor–critic update that treats action-independent terms as baselines (e.g., REINFORCE with baseline, A2C, PPO), the quantity* B_t *in* (35) *does not affect the expected policy gradient of agent* i *at time* t *because* B_t *does not depend on the realized action* $a_{i,t}$*. Consequently, adding* $r^{\mathrm{int}}_{i,t}$ *is* equivalent in gradient effect *to adding*

$$\widetilde{r}^{\mathrm{int}}_{i,t} \;:=\; \sum_{j\neq i} \widehat{\Phi}^{(j)}_m(o_t, a_{i,t}, a_{-i,t}) \tag{36}$$

to the reward, up to the standard baseline invariance.

Proof. Let $\pi_\theta(a_{i,t} \mid o_t)$ be agent i's policy and B_t be measurable w.r.t. $(o_t, a_{-i,t}, \text{exogenous})$ and independent of $a_{i,t}$. Baseline identity:

$$\mathbb{E}[\nabla_\theta \log \pi_\theta(a_{i,t} \mid o_t)\, C_t] = \mathbb{E}\left[C_t\, \nabla_\theta \sum_a \pi_\theta(a \mid o_t)\right] = \mathbb{E}[C_t\, \nabla_\theta 1] = 0, \tag{37}$$

for any C_t independent of $a_{i,t}$ given o_t.

Write $r^{\text{int}}_{i,t} = \widetilde{r}^{\text{int}}_{i,t} - B_t$ and denote by r_t the immediate reward used in the actor signal at time t. All standard actor–critic updates admit an actor gradient of the form

$$G_t(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_{i,t} \mid o_t)\, S_t]\,, \qquad S_t = \mathcal{L}_t[r] + Z_t, \tag{38}$$

where $\mathcal{L}_t$ is a linear functional of the reward sequence whose coefficient on the current r_t equals 1 and Z_t does not depend on r_t. Replacing r_t by $r_t - B_t$ yields

$$S'_t = \mathcal{L}_t[r - B] + Z_t = \mathcal{L}_t[r] - \mathcal{L}_t[B] + Z_t = S_t - B_t, \tag{39}$$

because the coefficient of the current reward in $\mathcal{L}_t$ is 1 and B_t pertains to time t only. Hence

$$G'_t(\theta) - G_t(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_{i,t} \mid o_t)\, (S'_t - S_t)] = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_{i,t} \mid o_t)\, (-B_t)] = 0, \tag{40}$$

by (37). Therefore the expected actor gradient is unchanged.

Instances of (38):

$$\text{REINFORCE with baseline: } S_t = G_t - b(o_t), \quad G_t = (\text{return, linear in } r_t), \tag{41}$$

$$\text{A2C/TD(0): } S_t = r_t + \gamma V(o_{t+1}) - V(o_t), \tag{42}$$

$$\text{GAE: } S_t = \sum_{\ell \geq 0} (\gamma\lambda)^\ell \delta_{t+\ell}, \quad \delta_t = r_t + \gamma V(o_{t+1}) - V(o_t). \tag{43}$$

In all cases the coefficient of r_t in S_t equals 1, hence the above argument applies.

For PPO with clipped surrogate

$$L_{\text{CLIP}}(\theta) = \mathbb{E}\Big[\min\left(r_t(\theta)\, \widehat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\, \widehat{A}_t\right)\Big], \quad r_t(\theta) = \frac{\pi_\theta(a_{i,t} \mid o_t)}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid o_t)}, \tag{44}$$

the gradient at $\theta = \theta_{\text{old}}$ satisfies

$$\nabla_\theta L_{\text{CLIP}}(\theta)\big|_{\theta=\theta_{\text{old}}} = \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a_{i,t} \mid o_t)\, \widehat{A}_t\Big]_{\theta=\theta_{\text{old}}}, \tag{45}$$

so replacing $\widehat{A}_t$ by $\widehat{A}_t - B_t$ changes the expected gradient by $\mathbb{E}[\nabla_\theta \log \pi_\theta(a_{i,t} \mid o_t)\, (-B_t)] = 0$ via (37).

Consequently, using $r^{\text{int}}_{i,t} = \widetilde{r}^{\text{int}}_{i,t} - B_t$ is equivalent, in expected actor gradient, to using $\widetilde{r}^{\text{int}}_{i,t}$.

By adding and subtracting the true others' return, write

$$\widetilde{r}_{i,t}^{\mathrm{int}} \;=\; \underbrace{\sum_{j\neq i} r^{(j)}(o_t, a_{i,t}, a_{-i,t})}_{\text{ideal prosocial term}} \;+\; \underbrace{\sum_{j\neq i} \big(\widehat{\Phi}_m^{(j)} - r^{(j)}\big)(o_t, a_{i,t}, a_{-i,t})}_{\delta_t},$$

where the residual δ_t satisfies $\big|\delta_t\big| \le (N-1)\varepsilon$ by the uniform approximation assumption. Thus, from the perspective of the policy gradient, using $r_{i,t}^{\mathrm{int}}$ is equivalent to augmenting the external reward by the *ideal prosocial term* plus a bounded residual.

Informal Proposition (Policy-gradient perturbation bound). Consider a γ-discounted actor–critic update for agent i with shaped return $R_i^{\mathrm{shp}} := \sum_{t\ge 0}\gamma^t\big(r_{i,t}^{\mathrm{ex}} + \alpha\, r_{i,t}^{\mathrm{int}}\big)$, where $\alpha \ge 0$. Let $\nabla J^{\mathrm{ideal}}(\theta_i)$ and $\nabla J^{\mathrm{mb}}(\theta_i)$ denote the expected policy gradients when the intrinsic term is respectively the ideal prosocial sum $\sum_{j\neq i} r^{(j)}$ and its model-based counterpart $\widetilde{r}_{i,t}^{\mathrm{int}}$ in (36). Assuming bounded rewards and standard mixing conditions for the Markov chain under policy π_θ, there exists a constant $C > 0$ (independent of ε) such that

$$\big\|\nabla J^{\mathrm{mb}}(\theta_i) - \nabla J^{\mathrm{ideal}}(\theta_i)\big\| \;\le\; C\cdot\frac{\alpha\,(N-1)\,\varepsilon}{(1-\gamma)^2}. \tag{46}$$

Sketch. The gradient difference reduces to a discounted sum of baseline-invariant per-step perturbations induced by replacing $\sum_{j\neq i} r^{(j)}$ with $\widetilde{r}_{i,t}^{\mathrm{int}}$, each bounded by $(N-1)\varepsilon$; accumulating over time yields the $(1-\gamma)^{-2}$ factor typical in PG perturbation analyses. △

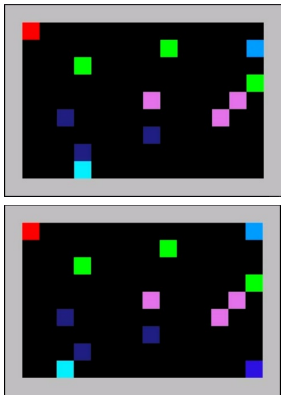
Equation (35) shows that $-\widehat{\mathrm{Regret}}$ behaves, in gradient effect, like adding a prosocial potential $\sum_{j\neq i}\widehat{\Phi}_m^{(j)}$ plus an action-independent baseline. When the proxy error ε and the shaping weight α are small, (46) suggests that training is close to the ideal prosocial shaping regime; conversely, large ε leads to a residual that can bias updates and should be mitigated via conservative α, improved modeling, or stronger regularization. If the maximization in (34) is approximated by K-candidate sampling or gradient ascent in continuous action spaces, the baseline B_t remains independent of the *realized* $a_{i,t}$ as long as the approximation procedure does not depend on $a_{i,t}$; hence the baseline-invariance argument still applies. Approximation quality only affects the tightness of the baseline and not the equivalence in (36).

B Additional details on experiments

B.1 Experiment Description

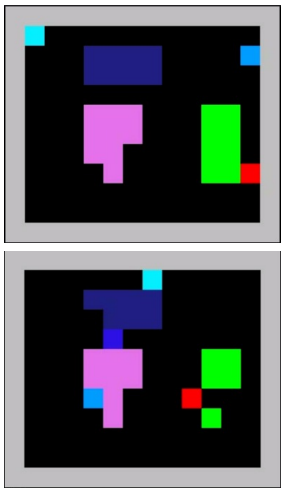
Level-based Foraging:

Agents are placed in the grid world, and each is assigned a random level. Food positions are determined in each episode, each having a level on its own(no more than 3). Agents can navigate the environment and can attempt to collect

[*LBF_3_Agents*]  [*LBF_4_Agents*]

food placed next to them. The collection of food is successful only if the sum of the levels of the agents involved in loading is equal to or higher than the level of the food. Finally, agents are awarded points equal to the level of the food they helped collect(two times if they are cooperating), divided by their contribution (their level).

Coin:

[*Coin_3_Agents*]  [*Coin_4_Agents*]

The reward for an individual agent in the environment at each time step under every scenario:

1. -4: other two agents get current agent's coin, while this agent does not get coin
2. -3: other two agents get current agent's coin, this agent gets a coin
3. -2: another agent get current agent's coin, this agent does not get coin

4. -1: another agent get current agent's coin, this agent gets a coin
5. 0: this agent do not get coin, other agents' do not get its coin
6. 1: this agent gets a coin

when the environment only contains two coins or one coin, the reward position of the missing reward would be (0,0), the type would also be (0). Let C_i be the coin type of agent i. $r_i(t)$ equals to the instantaneous reward of agent i at time step t. $S(t)$ equals to the set of all coin types in time step t.

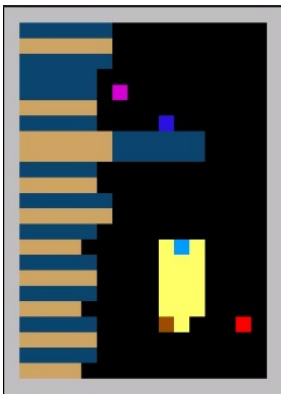
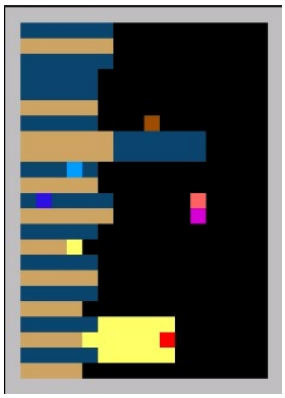
$$\delta_{C_i,T} = \begin{cases} 1 & C_i = T \\ 0 & otherwise \end{cases}$$

Therefore, the instantaneous reward of the agent i at time step t is:

$$r_i(t) = \sum_{T \in S(t)} \left(\delta_{C_i,T} - 2 \cdot \sum_{j \neq i} \delta_{C_j,T} \right)$$

In the four-agent setting of *Coin*, we introduce an adversarial agent by giving it a disruptive role. This agent has no matching coin type in the environment. Its primary function shifts to disturbing the dynamics of the game, potentially interfering with other agents' actions. The optimal prosocial policy for this modified agent would be to remain stationary and abstain from coin consumption, effectively minimizing its disruptive impact. This alteration creates a more complex strategic landscape, forcing the other three agents to adapt their behaviors in the presence of a potential adversary. The scenario now balances individual coin-collecting goals against the challenge of navigating an environment with an unpredictable, disruptive element, providing a richer context for studying multi-agent interactions and conflict resolution strategies.

Cleanup (Hughes et al., 2018):

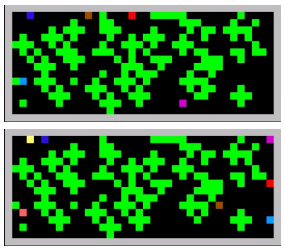
[*Cleanup_ 5*]  [*Cleanup_ 7*] 

In *Cleanup*, all agents are equipped with a fining beam which administers -1 reward to the user and -50 reward to the individual that is being fined.

There is no penalty to the user for unsuccessful fining. In Cleanup each agent is additionally equipped with a cleaning beam, which allows them to remove waste from the aquifer. Eating apples provides a reward of 1. There are no other extrinsic rewards. In Cleanup, waste is produced uniformly in the river with probability 0.5 on each timestep, until the river is saturated with waste, which happens when the waste covers 40% of the river. For a given saturation x of the river, apples spawn in the field with probability $0.125x$. Initially the river is saturated with waste, so some contribution to the public good is required for any agent to receive a reward.

We also provide the 7 agents edition for *Cleanup*. In the 7-agent edition of *Cleanup*, we expand the original environment to accommodate a larger group of participants, intensifying the complexity of social dynamics and resource management. The core mechanics remain unchanged: agents can clean waste from the river, collect apples that spawn based on river cleanliness, and use fining beams to penalize others. However, the increased number of agents creates a more crowded and competitive space, amplifying the tension between individual and collective interests. This expanded setting challenges agents to develop more sophisticated strategies for balancing personal reward maximization with the need for cooperative cleaning efforts. The larger group size also allows for the emergence of more complex social structures, such as temporary alliances or collective punishment of free-riders. Ultimately, this 7-agent version provides a more sophisticated experimental framework for investigating how prosocial behaviors and effective resource management strategies scale in larger multi-agent systems.

Common_Harvest (Hughes et al., 2018):

[*Common_ Harvest_ 5*]  [*Common_ Harvest_ 7*]

In *Common_ Harvest*, all agents are equipped with a fining beam which administers -1 reward to the user and -50 reward to the individual that is being fined. There is no penalty to the user for unsuccessful fining. Eating apples provides a reward of 1. There are no other extrinsic rewards.

In *Common_ Harvest*, apples spawn relative to the current number of other apples within an l^1 radius of 2. The spawn probabilities are 0, 0.005, 0.02, 0.05 for 0, 1, 2 and ≥ 3 apples inside the radius respectively. The initial distribution of apples creates a number of more or less precariously linked regions. Sustainable

policies must preferentially harvest denser regions, and avoid removing the important apples that link patches.

We also provide the 7-agents edition for the *Common_Harvest* environment. The 7-agent edition of *Common_Harvest* expands the original environment to create a more complex and challenging scenario for multi-agent cooperation and resource management. This version maintains the core mechanics of apple spawning based on local density and the use of fining beams, but introduces a larger group of agents competing for limited resources. The increased number of participants intensifies the challenge of maintaining sustainable harvesting practices, particularly in preserving the crucial links between apple patches. Agents must develop more sophisticated strategies to balance individual rewards with collective sustainability, navigating a more intricate social landscape where fining decisions and harvesting behaviors have broader implications. This expanded setting provides a richer platform for studying how sustainable resource management strategies scale with group size, the emergence of implicit social norms, and the potential for diverse role specialization among agents. Ultimately, the *Common_Harvest_7* offers deeper insights into complex multi-agent dynamics in shared resource scenarios, mirroring real-world challenges in environmental and economic systems.

In the *Common_Harvest* and *Cleanup*, agents use partially observed graphics observation, which contains a grid of 15×15 centered on themselves. Therefore, we could construct the environment as the POMG.

B.2 Algorithm Details

We utilized PPO algorithm in stable-baselines3 (Hill et al., 2018) to implement the baselines and our methods, with all the agents using separated policy parameters for every experiments. Following McKee et al. (2020), we implement Social Value Orientation (SVO) by modifying each agent's reward. The original formulation $r_i^{\text{SVO}} = r_i - \alpha(1 - \arctan(\sum_{j \neq i} r_j / r_i))$ is undefined when $r_i = 0$, which occurs frequently in our environments (e.g., when an agent does not collect a coin). To ensure numerical stability, we use a modified formulation:

$$r_i^{\text{SVO}} = r_i + \alpha \cdot \text{atan2}\left(\sum_{j \neq i} r_j,\, r_i + \epsilon\right), \tag{47}$$

where $\epsilon = 10^{-8}$ prevents division by zero, and $\text{atan2}(y, x)$ is the two-argument arctangent function that correctly handles all quadrants and the case $x = 0$. We set $\alpha = 0.5$ for all SVO experiments, corresponding to a prosocial preference angle of approximately 45 as recommended in McKee et al. (2020).

The hyper-parameters for PPO training are as follows.

- The learning rate is 1e-4
- The PPO clipping factor is 0.2.
- The value loss coefficient is 1.

- The entropy coefficient is 0.001.
- The γ is 0.99.
- The total environment step is 1e7
- The environment episode length is 1000.
- The grad clip is 40.

B.3 Detailed Implementation

Network Architectures. **Policy Network Φ^i_π:** For grid-world environments (Coin, LBF), we use a CNN encoder with three convolutional layers (32, 64, 64 filters, kernel size 3×3, stride 1, ReLU activations) followed by a GRU layer (hidden size 128) and two fully-connected layers (128 units, then $|\mathcal{A}|$ outputs). For pixel-observation environments (Cleanup, Common_Harvest), we use a deeper CNN encoder (four convolutional layers with 32, 64, 64, 64 filters) followed by the same GRU and fully-connected architecture.

Generative Model $\Phi^i_{\mathbf{m}}$: The generative model uses the same CNN encoder architecture as the policy but processes the concatenated joint observation (stacking observations from all agents along the channel dimension). The encoder output is concatenated with a one-hot encoding of the joint action and passed through a GRU layer (hidden size 256) followed by two fully-connected layers (256 units, then N outputs) to predict all agents' rewards.

Value Network: Shares the CNN encoder with the policy network, followed by separate fully-connected layers (128, 1) for value estimation.

Generative Model Training Configuration.

- Optimizer: Adam with learning rate 3×10^{-4}
- Batch size: 256 transitions sampled from replay buffer
- Update frequency: Every 1000 environment steps
- Loss function: Mean squared error (Eq. 5)
- Replay buffer size: 50000 transitions

Handling Partial Observability. Both policy and generative models use GRU layers to aggregate information from observation histories. Hidden states are reset at the beginning of each episode. During training, we use truncated backpropagation through time with sequence length 20. The GRU hidden state is stored in the replay buffer to enable proper credit assignment during off-policy updates of the generative model.

Counterfactual Maximization Implementation. For all environments, we compute the maximum in Eq. 4 by exhaustive enumeration over the discrete action space. The action space sizes are:

- Coin/LBF: 5 actions (4 cardinal directions + stay)
- Cleanup/Common_Harvest: 8 actions (4 directions + 2 rotations + fire/clean + stay)

This enumeration requires $|\mathcal{A}|$ forward passes through the generative model per agent per timestep, adding negligible computational overhead (less than 5% of total training time).

Hyperparameters for Main Experiments. Table 1 summarizes the key method-specific hyperparameters used in our experiments. All methods use the same PPO hyperparameters listed above.

Table 1: Method-specific hyperparameters for main experiments (Figures 2 and 3).

Environment	CF (α)	Inequity Aversion (β_1, β_2)	SVO (α)
Coin_3_Agents	2.0	(0.5, 0.5)	0.5
Coin_4_Agents	5.0	(0.5, 0.5)	0.5
LBF_3_Agents	2.0	(0.5, 0.5)	0.5
LBF_4_Agents	5.0	(0.5, 0.5)	0.5
Cleanup_5	2.0	(0.5, 0.5)	0.5
Cleanup_7	3.0	(0.5, 0.5)	0.5
Common_Harvest_5	2.0	(0.5, 0.5)	0.5
Common_Harvest_7	3.0	(0.5, 0.5)	0.5

Hyperparameter Selection. For our method (CF), we selected α based on preliminary experiments using the scaling heuristic $\alpha \approx N - 1$, where N is the number of agents. For baselines, we used default values from the original papers: Inequity Aversion uses $\beta_1 = \beta_2 = 0.5$ (equal weight on advantageous and disadvantageous inequity) (Hughes et al., 2018); SVO uses $\alpha = 0.5$ corresponding to a prosocial orientation (McKee et al., 2020). All methods received equal computational budget for training (2×10^7 environment steps) and used the same network architectures and PPO hyperparameters to ensure fair comparison.

B.4 Computational Resources

All experiments were conducted on an HPC system equipped with 128 Intel Xeon processors operating at a clock speed of 2.2 GHz and 40 gigabytes of memory.

4.2 Tripartite Tender Game Framework for MARL in Crowdsource First-Last Mile Logistics

Tripartite Tender Game Framework for MARL in Crowdsource First-Last Mile Logistics

Kondrashov I.[1][0009-0000-0306-2037], Batsanova E.[1][0009-0000-5939-4698],], Tomilov I.[1][0000-0003-1886-2867], Gusarova N.[1][0000-0002-1361-6037]and Vatian A.[1][0000-0002-5483-716X]

[1] ITMO University, Saint Petersburg 197101, Russia
international@itmo.ru

Abstract. This paper proposes a Multi-Agent Reinforcement Learning framework for crowdsourced First-Last Mile Logistics based on a tripartite tender game mecha-nism. The framework treats clients and couriers as active autonomous agents with diverse preferences while respecting labor regulations and operational constraints. We theoretically derive optimal bidding strategies using cumulative distribution functions and experimentally demonstrate convergence to optimal behavior in 2-agent and 10-agent scenarios. Results confirm positive social welfare and non-negative courier profits under feasible parameters, offering a scalable, citizen-centric solution for decentralized logistics coordination..

Keywords: First-Last Mile Logistics, Multi-Agent Reinforcement Learning, Auction Mechanisms, Game Theory

1 Introduction

First-Last Mile Logistics (FLML) is a type of supply chain widely used in cities and refers to transporting parcels from the customer to the sorting center (first-mile logistics) and vice-versa (last-mile logistics) [1]. It represents a critical socio-technical challenge where citizens are involved as both service recipients (clients) and employees (couriers) within platform economies. Therefore, unlike conventional supply chains, currently used FLML systems feature dual roles of citizens: (1) end-clients requiring time-sensitive delivery, and (2) crowdsource couriers, or regular agents, who function as citizens exercising autonomous decision-making during route execution. Crowdsource components in FLML may be represented as multi-agent system (MAS) [2]. Hence, it is appropriate to build a FLML system as a citizen-centric MAS (C-MAS) [3] coordinating the decentralized FLML participants with diverse needs and preferences. On the other hand, the FLML C-MAS must satisfy hard operational constraints protecting the interests of employees (for example, 12-hour shift limits) and customers (for instance, client's location stationarity) as well as soft constraints including the maximum tender price and the time window.

Today, most commonly applied MAS approaches tend to be hierarchically built and are commonly focused on upper-level market interactions, where clients and couriers are rarely considered as active elements of the system [4, 5, 6, 7. A step toward taking an employee's well-being into account, in particular, protection by local labor law, represent the approaches in which FLML couriers are active agents, for example, through auction mechanisms [1, 8, 9, 10]. However, [1] and [8] only focus on theoretically trivial games, rarely presented in the real FLML system, whereas [9] considers only soft constraints on the regular agents and [10] has not provided an analytical background of the auction system.

Our work develops an auction approach from the perspective of respecting the rights of all participants, including clients as active participants, thereby forming a novel scalable MARL approach based on cumulative distribution function (CDF)

optimization [11] and explicitly demonstrating practice-to-theory convergence. Behaving as autonomous agents in transaction processes, clients and couriers produce the local tripartite tender game, attempting to maximize each person's expected profit. Using modified notation from [12], tripartite tender game agents are (1) clients deciding the input parameters and choosing the winner, (2) couriers competing on prices inside the game and (3) Platform Decision Agent (PDA) which is a supportive algorithm to verify if the winner can meet the time window and maximum worktime conditions. This mechanism requires at least one client, initializing the tender procedure and anticipating a lower delivery cost than what they would pay for FLML to a large firm, and at least two active couriers, however, it can easily be scaled to an arbitrary number of participants. This framework provides cohesive theoretical background for the MARL-based explainable FLML systems. We propose the theoretical basis explaining the behavior of MARL agents, whose reward functions are based on the expected profit functions and basic rewards implemented to avoid rigid game cases. Appendix 1 presents a diagram of the framework implementing the tripartite tender game mechanism as well as a diagram of the global route execution process.

The proposed framework gives the non-strategic clients the possibility to architect the game by establishing the maximum tender price and by setting the time window that the courier must respect to successfully deliver orders. Also, the client decides whether to accept or reject the tender winner. On the other hand, couriers within the tender game are authorized to bid, set the lower price boundary and form unique policies to win the tender and thus make profit. The above procedure meets the needs of citizen as primary agents with diverse needs and preferences. Furthermore, the tripartite tender game mechanism provides positive social welfare in terms of the game theory [13,14] through non-linear price establishment.

Our contribution is as follows:

1. We adapted the game theoretical model [11] to the conditions of the citizen-oriented FLML scenario using hybrid constraints.
2. We theoretically derived the existence of nontrivial optimal strategies for the tripartite tender game for the case of 2 players, and experimentally demonstrated the convergence to the same optimal behavior for the case of the K number of players.
3. We experimentally showed positive client's welfare and non-negative profits for the couriers participating in the tender, as well as the convergence of the game to the optimal strategy in the case of mean-field approximation.

2 Methods and materials

2.1 Theoretical basis of the work

Game parameters within the framework are as follows:

$B = const$ refers to couriers hard constraint of upper boundary of route execution time meeting labor protection conditions; $G = (N, P, \Pi)$ is game structure, where: $N = (i, j)$ is the set of agents; $P = (P_i, P_j)$ is the set of agents' actions; $F_i(P_i)$ is the cumulative distribution function of agent i for tender price P_i. $\Pi = (\Pi_i(P_i|P_j), \Pi_j(P_j|P_i))$ is the set of profit functions; br_i is the basic reward of agent i; δ is the probability of agent i winning the tender under the tie situation $P_i = P_j$; $p(P_i < P_j)$ is the probability of agent i's tender price being less than agent j's tender price; T_i^{est} is the estimated time of route execution for agent i, $T_i^{est} \leq B \ \forall i \in N$; T_i^{real} is the real route execution time of agent i; U is the citizen-driven soft constraint as the maximum tender price will be accepted; V is the winner's reward share such that $br_i + V + br_j = 1$ (the platform takes a commission from winner's total reward), $\epsilon = const$, $\epsilon \ll 0.1$; R_i, R_{i*} are the routes of agent i under the basic and modified optimal route execution; A is the set of orders to deliver; $TW(A)$ refers to client's soft constraint as the time window when A must be delivered; $f_i(TW(A))$ is the cost of the A incorporated into the route R_i for agent i.

We introduce the tender profit function of agent i as

$$\Pi_i(P_i) = br_i P_i + P_i\,(p(P_j > P_i) + \delta(P_j = P_i)) - f_i(TW(A)) \quad (1)$$

Lemma 1 (Parameter Validity and Equilibrium Existence). Assumption: under boundary hard constraints $T_{real} \leq B$ and soft constraints $br_i > br_j > 0$, $br_i + br_j + V = 1,\ U > 0, f_i\big(TW(A)\big) = f_j\big(TW(A)\big) = 0\ \forall i \in$ N and under the CDF equilibria concept

$$\Pi_i\Big(P\Big|F_i^*(P), F_j^*(P)\Big) \geq \Pi_i\left(P\Big|F_i(P), F_j^*(P)\right)\ \forall F_i(P) \neq F_i^*(P);\ i \neq j;\ i, j \in N \quad (2)$$ the following propositions satisfied:

- The lower price bound $P_i^{lower} > 0$
- A mixed-strategy Nash equilibrium exists with CDF $F_i^*(P)$ satisfying the indifference condition
- The equilibrium is unique and stable under the given constraints □

Lemma 1 shows that for 2-agent tender game under soft client-driven $(U, TW^{(A)})$, couriers-driven $(br_i, br_j, f_i{}^{U}{}_{,j})$ and hard (B) constraints the equilibria solution (F_i^*, F_j^*) is the respective agent-determined tender price CDF curves, which are optimal as long as no parameters are shifted. Since the CDF curve is analytically formed, the RL agent's behavior may be explained according to the theoretical concept, which we adapted for the proposed scenario.

*We assume the case without additional platform payoff, and the platform's purpose is exactly maintaining tender games.

To secure the transparency and feasibility of the tender game results, we add the route feasibility conditions to the game terms. Let $R_i \to R_i^*$ be the adaptation $O(R_i|A,TW^{(A)})$ transition of courier i's route under the specific optimization policy and $T_i(R_i^*)$ the total route execution time. The tender game implies 2 validation layers: the agent i's selfvalidation and the tender owner's winner validation. From the self-validation perspective, the $f_i{}^{U}{}_{,j}$ value has been added to address the negative expected profit from winning the tender. The second validation implies the indifferent review procedure on the possibility of the tender winner to complete A.

The winner must share his full route, the total time cost and the time window restrictions for A with the Platform Decision Agent (PDA), the PDA calculates the possibility and transfers it to the owner, then the owner accepts/rejects the winner candidate. If a winner w has been rejected, $\Pi_w([T_w(R_w^*|A) > B] \vee (R_w^* = \emptyset)) = -br_w$, and the tender owner replays it once again for $N' \equiv N \backslash \{w\}$. All agents in N' receive a new offer and the sets of transparent information and act according to the standard protocol, otherwise the tender owner declares the winner and their bid, and the winner executes new route R_{w*}. This additional approval mechanism ensures that (a) bids remain undisclosed before the end of the game, (b) the winner's full route is not disclosed to the tender owner (privacy security), (c) the tender owner gets proved and trustworthy information about the winner's possibility of executing the modified optimal route and (d) the courier gets the new optimal route to execute it as fast as possible to meet the time windows requirements.

Using the above transactions mechanism, we can construct a sequence of tender games for which the social welfare conditions are satisfied in accordance with Theorem 1 (the proof is given in Appendix 3).

Theorem 1: Assumption: if for each tender game in the sequence $\mathbb{G}^{\tau} = G_{N_1}^{\tau} \oplus \ldots \oplus G_{N^{\tau_l}}$ the boundary constraints from Lemma 1 are met, the social welfare function value cannot be negative for any feasible set of parameters in sequential decentralized tender games system:

$$\mathbb{W} = \sum_{\tau=1}^{T} \mathbb{W}^{\tau} \geq 0 \quad (3)$$

Theorem 1 proves that negative social welfare results can never occur in client-driven local tender games thus, both crowdsource couriers and clients achieve positive social welfare from decentralized local tender games sequence.

2.2 Experiment design

The experimental MARL system was based on the following conditions:

1. For RL agents trained to play tender game according to Lemma 1 we used Centralized Training with Decentralized Execution (CTDE) learning mechanism discussed in [15] to avoid non-stationarity problem within learning iterations. According to Yiqin W. [16, par.3.1.3], "Orthogonally, …, to tackle the non-stationary issue, the centralized critic training decentralized execution (CTDE) paradigm **(S3)** has great significance. …. Under this architecture [CTDE], since agents do not experience unexpected changes in the dynamics of the environment, the training procedure and obtained results can be stabilized."
2. Learning reward function was normalized to 1 without any loss of representability.
3. MARL factorial cost of calculating K 1-to-1 tender games in K-agents structures was addressed according to Yaodong Y. et.al. Authors state, "The scalability issue of multiagent learning in non-cooperative general-sum games can also be alleviated [from the MARL perspective] by applying the mean-field approximation directly to each agent's Q-function" [17, ch.9.3], supporting the mean-field approximation as the viable MARL complexity reduction approach.
4. This MARL system implicitly works with different couriers' logistics costs via P_i^{lower} analytical value as the baseline profit value under the indifference condition. Thus, the cost differences robustness is implemented implicitly into the MARL system without additional code necessity.

MARL description:

Observation of all basic components of the MARL system is presented in Table 1.

Component	**Description**
Algorithm	CTDE Actor-Critic, Beta policies, TD(0) critic
Actor Net	4→128→128→[α,β], ReLU, softplus output
Critic Net	(4+1)→128→128→128→1, ReLU
State	[P_lower, U, br_i, br_j, (br_i≥br_j)]
Action	a~Beta(α,β), bid=P_lower+a·(U-P_lower)
Reward	Normalized expected profit + basic reward for losses For exact formulation look see Appendix 2, A1
Optimizer	Adam (actor: 1e-4, critic: 1e-3), grad clip=1.0
Hyperparams	Inflation rate (γ)=0.99, Target update (τ)=0.005, batch size=64, Replay buffer=10^6, Temperature (T)=2.0, Entropy coefficient (α_ent)=0.05
Training	100 epochs × 128 tenders, update every 16 steps
Seeds	{42, 123, 456, 789, 1011} for statistical reporting

Table 1. Representation of the MARL system components.

To construct the practical CDF equilibria concept, we used Beta distribution-based agents with initial state $B(1,2)$ and the learning procedure presented in Figure 1.

```
# Learning procedure pseudocode:
for each tender t:
    s_t ← get_state(br_i, br_j, P_lower, U)      a_t ~ Beta(α_θ(s_t)/T,
β_θ(s_t)/T)          # Temperature-scaled sampling      bid_t ← P_lower +
a_t·(U - P_lower)
    r_t ← compute_reward(bid_t, won, br, F_opp)  # Normalized profit
    store_transition(s_t, a_t, r_t, s_{t+1})

    if buffer_size ≥ BATCH_SIZE:
        Q_target ← r + γ·(1-d)·Q_θ'(s', mean[π_θ'(s')])
L_critic ← MSE(Q_θ(s,a), Q_target)          θ ← θ -
α_c·∇L_critic
        A ← normalize(Q_θ(s,π_θ(s))) - baseline
L_actor ← -E[log π_θ(a|s)·A] - α_ent·H[π_θ]          θ
← θ - α_a·∇L_actor
        θ' ← τ·θ + (1-τ)·θ'
```

Fig. 1. Learning procedure pseudocode.

Based on the initial suggestion that Beta distribution-based agent's policy will tend to the theoretically proven optimal policy, we proposed the agent and critic network structures presented in Figure 2.

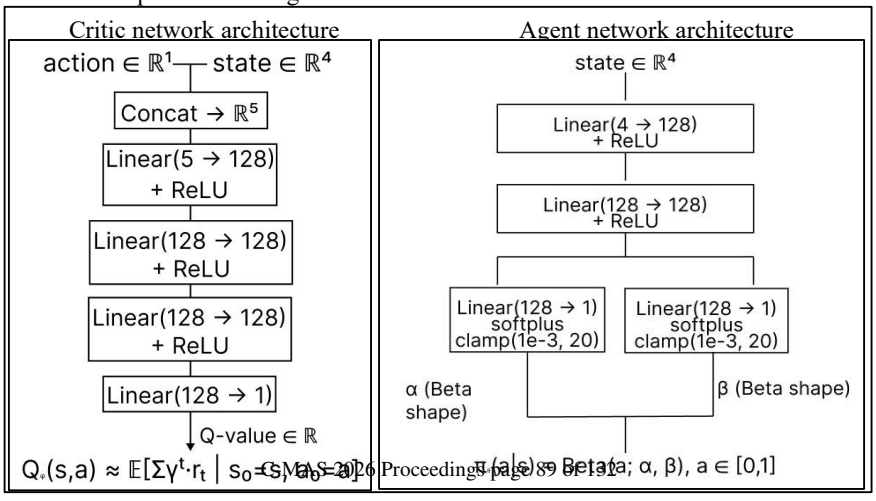


Fig. 2. Agent (left) and critic (right) Neural Network schemas.

To evaluate the models' performance, we used double-validation logic common for CTDE learning framework. Firstly, the divergence between the optimal expected profit and empirical expected profit is calculated by critic network which has the full observation of the game state. After that, each agent network's action is examined via the critic network as the MSE between

the optimal parameters of policy and the empirical agent's policy at the evaluation step to evaluate the changes in current policy. Beside it, agent also observes empirical profits to support the policy optimization gradient. Action space and episodes summary are provided in Table 2 and Table 3 respectively.

Property	Specification
Type	Continuous, bounded
Normalized action	a ∈ [0, 1] (sampled from Beta distribution)
Physical bid	$bi = P^{l} + (U - P) \in [P, U]$
Policy parameterization	$(\mid s) = B\,t\,(; (s), (s))$
Training exploration	Temperature scaling: Beta(α/T, β/T) with T = 2.0
Execution (eval)	Deterministic: a = α / (α + β) (Beta mean)

Table 2. Action space summary.

Parameter	Value	Description
Epochs	100 (configurable)	Outer training loop
Tenders/epoch	128 (configurable)	Environment interactions per epoch
Total interactions	~12,800 tenders	Typical training budget
Evaluation	Every epoch	KS-test convergence + win-rate tracking
Convergence criterion	KS p-value > 0.8	Empirical CDF matches theoretical (Lemma 1)

Table 3. Episodes and evaluation summary.

We conducted three experiments. The goal of the first experiment was to demonstrate the convergence of the empirically obtained CDF to the theoretical one in games with different numbers of agents. The agents' objective functions and theoretical CDFs were taken from Lemma 1. To analyze the convergence of the empirically obtained CDF to the theoretical one in the 2-agent game, we used the Kolmogorov-Smirnov test. For the 10-agent game, we used a mean-field approximation of the opponent's response function for each agent i, recorded the p-values for the last 40 agent training epochs, and plotted a boxplot of the p-values.

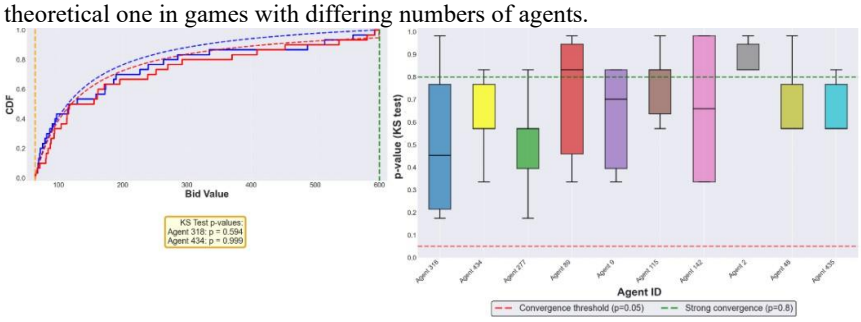
In the second experiment, we examined the relationship between the agent CDFs and the bid size in a 2-agent game with different input parameters. The client's welfare and the winner's profit were calculated according to Theorem 1. The agent's objective function and the theoretical CDFs are given in Lemma 1. For the 2-agent game, we first set a low-cost time window and a high maximum tender price followed by shifting the parameters to a fixed, busy time window and a low maximum tender price.

The third experiment investigated the scalability of the system and the dependence of the CDF within a 10-agent system on the bid size under different game initialization parameters. The agent's objective function for the profit was taken from Lemma 1, the theoretical CDF was calculated as the average CDF of mean-field approximations across agents, and the client's welfare and the winner's profit were calculated according to Theorem 1. A visual representation of the convergence of the experimental CDFs to the theoretical mean used the maximum positive and the maximum absolute negative deviations of the experimental CDF from the theoretical CDF for each bid size. For the 10-agent game, we first used a low-cost time window and a maximum tender price followed by the selection of a fixed, busy time window and a low maximum tender price.

3 Results and discussion

3.1 Experiments result under the normal state environment and proposed agents architecture

Figure 2 shows the results of evaluating the convergence of the practical CDF to the theoretical one in games with differing numbers of agents.

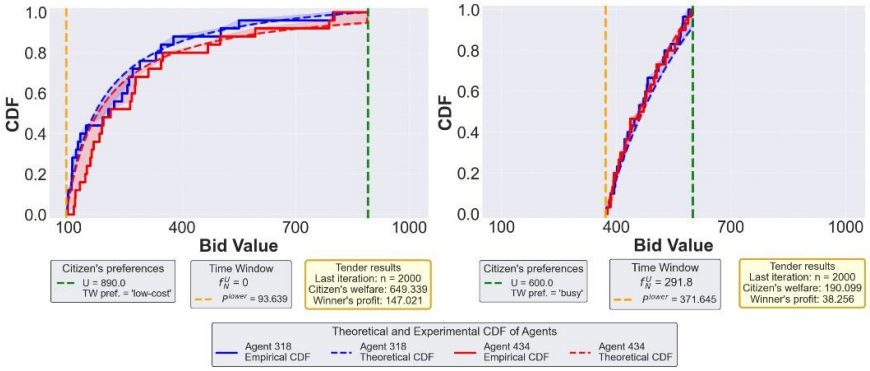


a b

Fig. 2. Convergence of empirically obtained CDFs of MARL agents to the theoretical optimal CDF under the conditions of high marginal cost and low-cost time window: a - for the 2-agent game, b - for the last 40 iterations of the 10-agent game.

Figure 2a shows that the RL pricing distribution tends to the precalculated theoretical CDFs. Figure 2b demonstrates that the p-value of the significance of the discrepancy between the empirically obtained CDF and the theoretical mean-field approximated CDF distribution satisfies the conditions of the Kolmogorov-Smirnov test.

Figure 3 shows the dependence of the agent CDF on the bid amount under the constraints on the tender price and the client's preferred time window for a 2-agent game.

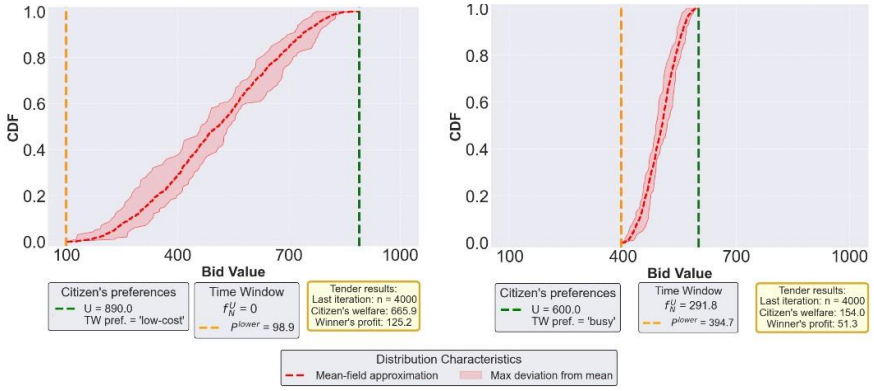


a b

Fig. 3. CDF dynamics for a 2-agent game under the different citizen-defined parameters: a - for a high maximum tender price U in a low-cost time window, b - for a low maximum tender price U and a busy time window.

Figure 3 shows that if the citizen's parameters $TW(A), U$ and the couriers' constraints br_i, br_j, $f_i{}^U{}_j$ meet the feasibility bounds from Lemma 1, both the citizen's and the winner's welfare functions are positive. It also demonstrates that both agents tend to adhere to the theoretical optimal CDF for different input parameters of the game.

Figure 4 shows the mean CDF patterns in a 10-agent MARL system and the client's welfare and the mean CDF shifts in tighter citizen-driven soft constraints.



a b

Fig. 4. Dynamics of approximate theoretical CDF and empirically obtained deviation fields in the 10-agent game under the different client-driven parameters: a - for high maximum tender price U in a low-cost time window, b -for low maximum tender price U and a busy time window.

Figure 4 demonstrates that under the different citizen-driven and couriers' soft constraints the positive social welfare result remains while the feasibility condition from Lemma 1 has been met. Also, it shows that there is no tendency of abnormal behavior or unpredictable patterns within 10-agents MARL.

3.2 Ablation

In this subsection we provide the results of the experiments under the following patterns ablation:

1) Set br_j as 0 value to examine the optimal behavior shift within both agents if one agent has no basic reward
2) Set $B = 1$ to relax the learning aggressiveness constraint
3) Set $br_i = br_j = 0.5$, thus reducing the tender expected benefit to 0 for any $i \in N$
4) Finally, we set $br_i = br_j = 0$ to construct a rigid reverse auction situation

Figure 5 demonstrates the tendency of agent with higher basic reward ratio to excessively prefer the high-bid strategy over the theoretically estimated strategy while the lower basic reward agent's empirical policy difference from the theoretical optimal policy tends to zero.

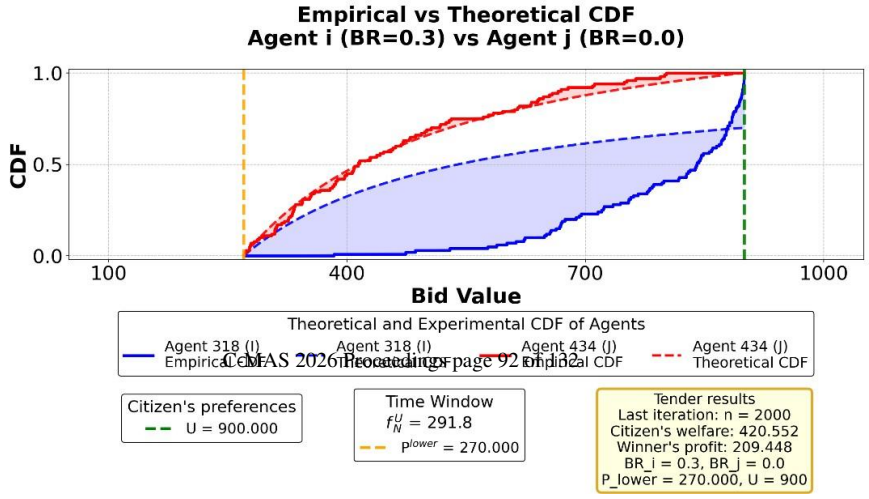


Fig. 5. Empirical CDF comparison with theoretically optimal strategy under the proposition of lower basic reward is zero while the highest basic reward remains.

Figure 6 shows the tendency of higher basic reward agent to hedge the risks and bidding lower on average than is expected by the analytical solution if the aggressiveness constraint is relaxed.

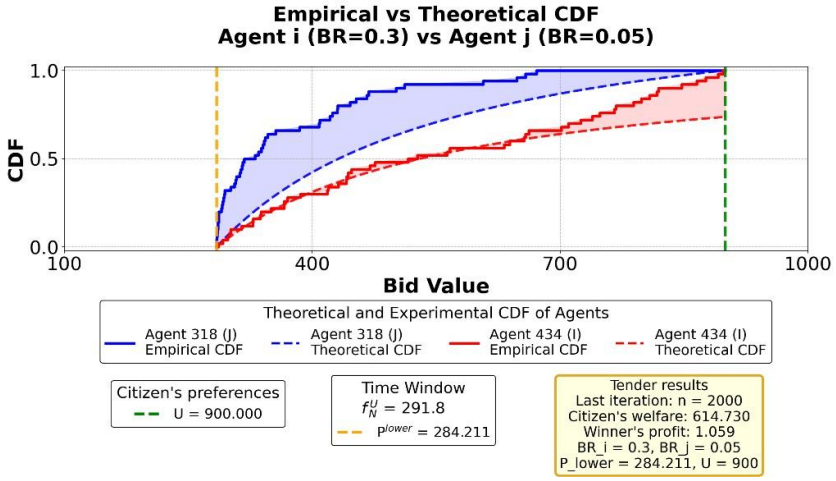


Fig. 6. Empirical CDF comparison with theoretically optimal strategy under the relaxed learning aggressiveness constraint.

Figure 7 demonstrates the rigid game case where no strategy space is involved ($br_i + br_j = 1 \rightarrow V = 0 \Leftrightarrow P^{lower} = U$).

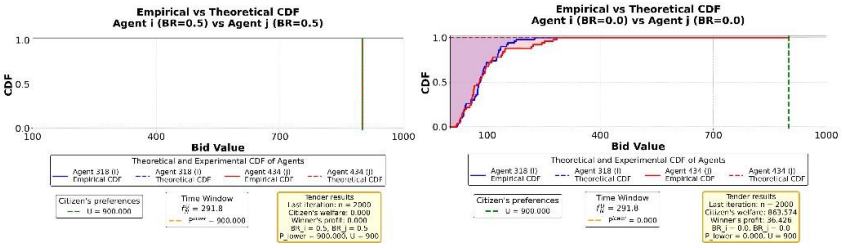


Fig. 7. Empirical CDF comparison with theoretically optimal strategy under the two rigid cases: the zero-strategy space and zero basic reward case of Bertrand rigid game model.

3.3 Baseline comparison

In this section we demonstrate the empirical performance of agents described in section 2.2 against the baseline bidding agents. To compare the tender game result, we used two typical baseline bidders: fixed-bid strategy agent and heuristic bidder. Fixedbid strategy agent is a common agent, bidding exact price indifferently to the opponent's decisions and game history. Heuristic bidder's architecture was built according to the Heymann and Mertikopoulos [18], incorporating the sufficiently large stochastic bidding space and starting with uniform distribution CDF. Heuristic agent learning pseudocode presented in Appendix 4.

Figure 8 shows the last 100 iterations performance of the MARL and fixed-price agents inside the tender game with equal parameters of basic rewards.

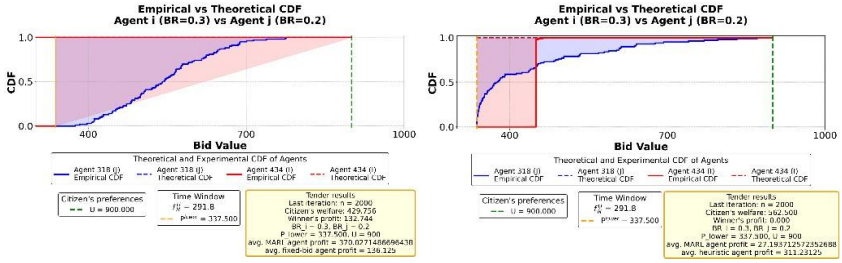


Fig. 8. Empirical profit comparison between the MARL and fixed-bid agents within 2000 iteration training and 100 samples (left) and empirical profit comparison between the MARL and heuristic agents within 2000 iteration training and 100 samples (right).

4 Conclusion

This paper proposes a MARL-based framework for organizing Crowdsource FirstLast Mile Logistics, which coordinates the decentralized FLML participants as active agents with diverse needs and preferences while taking into account the regulatory constraints. Based on the findings in [11] and by introducing hybrid restrictions, we adapted the FLML scenario and theoretically demonstrated the nontrivial optimal strategies appropriate for the tripartite tender game. We experimentally confirmed the convergence of a two-agent game with MARL to theoretically optimal CDFs and the robustness of this MARL behavior for the 10-player case.

The proposed framework can be easily integrated into existing crowdsourcing platforms. The framework code is available on request.

In this work we intentionally excluded the logistics and market modelling due to the volume limitations, relevant dataset absence and scope of the article. Delivery time window robustness, non-stationary demand and exception handling via automatic logistics systems like PDA will be presented in future works with the extended analytical model to larger K-agent games, incorporate stochastic travel-time distributions, and evaluate performance on real-world FLML datasets.

Acknowledgments. This work supported by the Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002), agreement No139-15-2025-010.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

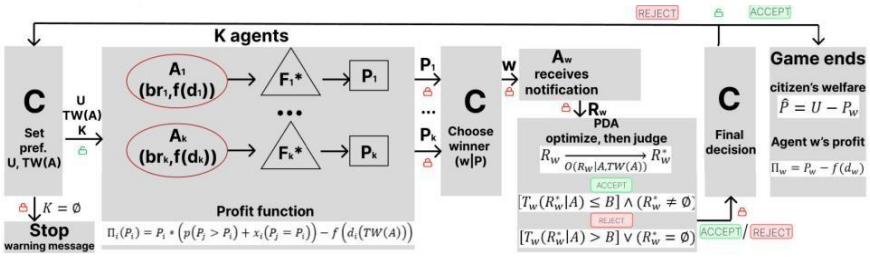
References

1. Y. Li, Y. Li, Y. Peng, X. Fu, J. Xu and M. Xu, "Auction-Based Crowdsourced First and Last Mile Logistics," in *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 180-193, Jan. 2024, doi: 10.1109/TMC.2022.3219881.
2. Rebollo, M., Giret, A., Carrascosa, C., Julian, V. (2018). The Multi-agent Layer of CALMeD SURF. In: Belardinelli, F., Argente, E. (eds) Multi-Agent Systems and Agreement Technologies. EUMAS AT 2017 2017. Lecture Notes in Computer Science(), vol 10767. Springer, Cham. https://doi.org/10.1007/978-3-030-01713-2_31
3. Sebastian Stein and Vahid Yazdanpanah. 2023. Citizen-Centric Multiagent Systems: Blue Sky Ideas Track. In Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023, IFAAMAS, 6 pages. Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
4. Silva, M.; Pedroso, J.P.; Viana, A. Deep reinforcement learning for stochastic last-mile delivery with crowdshipping. *EURO J. Transp. Logist.* **2023**, *12*, 100105.

5. Russo, F.; Comi, A. Urban Courier Delivery in a Smart City: The User Learning Process of Travel Costs Enhanced by Emerging Technologies. *Sustainability* **2023**, *15*, 6253.
6. Wang, L.; Xu, M.; Qin, H. Joint optimization of parcel allocation and crowd routing for crowdsourced last-mile delivery. *Transp. Res. Part B Methodol.* **2023**, *171*, 111–135.
7. S. D. Handoko, D. T. Nguyen and H. C. Lau, "An auction mechanism for the last-mile deliveries via urban consolidation centre," *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, New Taipei, Taiwan, 2014, pp. 607-612, doi: 10.1109/CoASE.2014.6899390.
8. Li, Yafei & Li, Yifei & Peng, Yun & fu, Xiaoyi & Xu, Jianliang & Xu, Mingliang. (2022). Auction-Based Crowdsourced First and Last Mile Logistics. IEEE Transactions on Mobile Computing. PP. 1-13. 10.1109/TMC.2022.3219881.
9. Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, Weinan Zhang. 2018. Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising. In 2018 ACM International Conference on Information and Knowledge Management (CIKM '18), October, 2018, Torino, Italy
10. Zhang, Jincheng. (2025). MAS-Based Adaptive Cooperative Optimization Algorithm (MAS-ACOA). 10.5281/zenodo.17853096.
11. Narasimhan, Chakravarthi. "Competitive Promotional Strategies." *The Journal of Business* 61, no. 4 (1988): 427–49. http://www.jstor.org/stable/2352790.
12. Lou, Zhenkai & Hou, Fujun & Lou, Xuming & Zhai, Yubing. (2021). Tripartite game models in a dual-channel supply chain: Competition and cooperation. RAIRO - Operations Research. 55. 10.1051/ro/2021029.
13. Nguyen, Van & Huynh, Vuong & Duong, Ho & Bui, Huu & Ha, Hai & Le, Quang & Ngo, Le & Nguyen, Tan & Nguyen, Ngoc & Nguyen, Hoai & Song, Zhao & Trang, Le & Han, The Anh. (2025). Social welfare optimisation in well-mixed and structured populations. 10.48550/arXiv.2512.07453.
14. Geffner, Ivan & Oesterheld, Caspar & Conitzer, Vincent. (2025). Maximizing Social Welfare with Side Payments. 10.48550/arXiv.2508.07147.
15. Zhao J, Hu X, Yang M et al (2022b) Ctds: centralized teacher with decentralized student for multiagent reinforcement learning. IEEE Trans Games 16(1):140–150.
16. Wang, Y., Wang, Y., Tian, F. *et al.* Intelligent games meeting with multi-agent deep reinforcement learning: a comprehensive review. *Artif Intell Rev* **58**, 165 (2025). https://doi.org/10.1007/s10462-025-11166-1.
17. Yang Y., Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective //arXiv preprint arXiv:2011.00583. – 2020.
18. Heymann, B. and Mertikopoulos, P., "A heuristic for estimating Nash equilibria in first-price auctions with correlated values", arXiv e-prints, Art. no. arXiv:2108.04506, 2021. doi:10.48550/arXiv.2108.04506.

Appendices

Appendix 1. Graphic illustration of the proposed tripartite tender game framework



C-MAS 2026 Proceedings page 95 of 132

Fig. A1. Tender game scheme

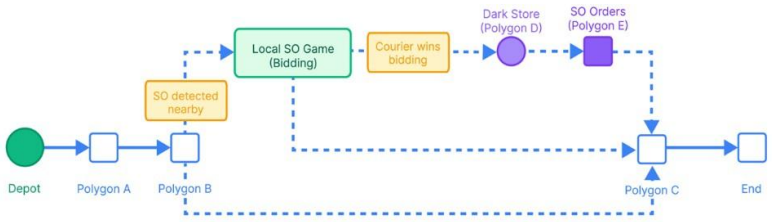


Fig. A2. Global route execution process, including the tender game structure

Appendix 2. Proof of Lemma 1

Agent i faces 3 opportunities:

$$\Pi_i(P_i) = \begin{cases} br_i P + VP_i, P_j > P_i, \\ br_i P + \delta VP_i, P_j = P_i, \\ br_i * P_i \quad , s \end{cases} \quad (A1)$$

Let us accept without losing generality that $br_i \geq br_j$.

The lower price bound is defined by couriers and implemented as:

$$P_{lower} = m\,x \left(\frac{br_i * U + f_i(TW(A))}{br_i + V}; \frac{br_j * U + f_j(TW(A))}{br_j + V} \right)_i$$

defines the minimum price of tender for the better performing agent to achieve the KPI-expected level and secure strongly positive profit. Set $f_i{}^U{}_j = \max\left(f_i((TW(A)); f_j(TW(A))\right)$ – the maximal expected cost to deliver orders set A exactly in delivery time window $TW^{(A)}$.

$$m\,x_{F_i(P)}\, \mathbb{E}(\Pi_i) = \int_{Pilower}^{U} \{br_i P + [1 - F_j(P)VP]\} dF_i(P) \quad (A2)$$

since expected profit of i cannot be less than $br_i U + f_i{}^U{}_j, w\ \ propos$:

$$br_i P + [1 - F_j(P)]VP = \begin{cases} {}^U, br_i U + f_{iU,j} P \\ br_i U + f_{i,j} \quad \leq < U \\ br_i + V \end{cases} \quad (A3)$$

$$br_j P + [1 - F_i(P)]VP = (br_j + V) * \begin{cases} \frac{br_i U + f_{iU,j}}{{}_i +}, if\ \frac{br_i U br_i U + f_{iU,j}}{br_i + V} \leq P < U \\ br_{(P)} = 1 - \frac{br_i (U - P) f_{i,j}^U}{V} \end{cases} \quad (A4)$$

$$[1 - F_j(P)] = \frac{br_i U + f_{i,j}^U - br_i P}{VP} \rightarrow F_j \quad VP \quad (A5)$$

$F_j(P_{lower}) = 0, F_j(U) = 1$

This proposition follows from 2 logical assumptions when the conditions of Lemma 1 are met:

1) j will never bid on $P' < P^{lower}\ sinc\ br_i P' + VP' < br_i P^{lower} + VP_{lower} \Rightarrow \Pi_i(P') < \Pi_i(P_{lower})\ \ n$

2)j will never bid on $U+\epsilon$ since $\Pi_i(U+\epsilon)=0<br_iU$ Thus $F_j(U)=1$ and $F_i(P)$ has a mass point in U. Since

$$F_j(P)=1-\frac{br_i(U-P)+f_{iU,j}}{VP\ br_iU} \quad (A6)$$

$$1-F_i(P)=\frac{\left(br_j+V\right)*\frac{+f_{iU,j}}{br_i+V}-j}{VP} \quad (A7)$$

$$=1+\frac{j}{V}-\frac{F_i(P)}{(br_i+V)VP}\left(br_jP+f^U\right)\left(br_j+V\right) \quad (A8)$$

Then

$$F_{i*}=\begin{cases}0, P<P_{lower},\\ 1+\frac{j}{V}-\frac{F_i(P)}{(br_i+V)VP}\left(br_jP+f^U\right)\left(br_j+V\right), P_{lower}\\ \leq P\leq U,\\ 1, P\geq U\end{cases} \quad (A9.1)$$

$$F_j^*=\begin{cases}1-\frac{\left(br_i+f_{i,j}^U\right)(U-P)}{VP}, P_{lower}\leq P\leq U\\ 1, P\geq U\end{cases} \quad (A9.2)$$

And

$$(\Pi_k^*)=\int_{P_{lower}}^{U}\mathbb{E}\{br_kP+[1-F_{-k}^*(P)VP]\}dF_k^*(P) \quad (A10)$$

expected tender profit for agent $k,\forall k\in N$.■

Appendix 3. Proof of Theorem 1

Assume the citizen's welfare for game G denoted as

$$P=U-\min(P_i,P_j) \quad (A11)$$

Then the social welfare function is:

$$W(G)=\hat{P}+\Pi_w(P_w|P_{-w})+br_{-w}P, \quad (A12)$$

where w – winner agent in tender game G.

$$\forall G(N,P,\Pi), br_i\geq 0\ \forall i\in N, \quad (A13)$$

$$P\geq 0 since\ \min(P_i,P_j)\leq U, \quad (A14)$$

$$n\ \Pi_w(P_w|P_{-w}, br_w=br_{-w}\leq 0\ or\ U\leq 0)=0 \quad (A15.1)\ \Pi_w$$

$$(P_w|P_{-w},[br_w>0\vee br_{-w}>0]\wedge U>0)>0 \quad (A15.2)$$

Where 15.1 represents rigid game structure and 16.2 represents normal tender game. Since $P\geq 0\ (15), br_i\geq 0\ \forall i\in N(14)\ \ n\ \ \Pi_w\geq 0(16.1;16.2), W(G)\geq 0$ – social feasibility is always non-negative under the parameters described in Lemma 1.

Global FLML process may be described through the following model: let $G_{N^\tau}(N,P,\Pi)-$ tender game with agents subset N, $|N|\geq 2, N\subset C$ and $\xi_{i^\tau,GN^\tau}\in[0;1]-$ binary classification of agent i participation in local tender game G_{N^τ}, $\kappa_{i^\tau,GN^\tau}\in$

$[0; 1]$ – binary classification of agent i winning local tender game G_{N^τ}. For more than 1 tender game processed within iteration τ, $N_a \cap N_b = \emptyset$. Incorporating the feasibility conditions 17.1-17.4,

$$\sum_{i \in N}^{N} \xi_{i^\tau, GN\tau} = |N|, \quad (A16.1)$$

$$\sum_{\in \subset i}^{C} \xi_{i^\tau, GN\tau} \leq |C|, \quad (A16.2)$$

$$\sum_{\in i N}^{C\ N} \kappa_{i^\tau, GN\tau} = 1, \quad (A16.3)$$

$$\sum_{\in \subset iN, NC}^{C} \kappa_{i, GN\tau}^{\tau} \leq \frac{|C|}{2} \quad (A16.4)$$

We define $\mathbb{G}^\tau = G_{N_1}^\tau \oplus \ldots \oplus G_{N_l}^\tau$ – full set of tender games at iteration τ.

Thus,

$$\mathbb{W}_\tau = \sum_{i=1}^{l} W\left(G_{N_i}^\tau\right) = \sum_{i=1}^{l} \sum_{k=1}^{N} \left(\hat{P}_i + \kappa_{k\tau, GN\tau} * \Pi_k(P_k | P^{-k}) + \sum br_{-k} P\right)_{or}$$

$$\mathbb{W}_\tau = \sum_{i=1}^{l} \left(\hat{P}_i\right) + \kappa_{k\tau, GN\tau} * \Pi k(P_k | P{-k}) + \xi_{m\tau, GN\tau} br_m P_m,$$

$$\left(\{k\} \in {}_{K | \kappa_{k^\tau, GN\tau}} = 1\right), m \in C \backslash \{K\} \quad (A17)$$

Finally,

$\mathbb{W} = \sum_{\tau=1}^{T} \mathbb{W}^\tau \geq 0$ $(A18)$ is a full social welfare from tender games within the observation period T. ■

Appendix 4. Heuristic bidder learning pseudocode

Input parameters: agent id - unique identifier
of agent P_lower, U - lower and upper bounds of bid
lr - learning rate
K - learning history window size
N - number of discrete bid action space
$\mathcal{A}$ = {a1,...,aN} - action space π =
{π1,...,πN} - policy vector
$\mathcal{H}$_bid - history buffer
opp_bids - history of opponent's bids within the K observations

Algorithm 1: Act
Input: state (unused), training ∈ {True, False}
Output: (action_val, action_tensor, log_prob_tensor)

```
if |ℋ_bid| < K then
chosen_idx ← UniformRandom({1, ..., N})
else
chosen_idx ← CategoricalSample(π)
raw_bid ← 𝒜[chosen_idx]
bid ← Clip(raw_bid, P_lower_init, U_init)
if U_init == P_lower_init then
action_val ← 0.5
else
action_val ← (bid - P_lower_init) / (U_init - P_lower_init)
end if
prob ← π[chosen_idx]
log_prob ← log(prob + ε)
Record bid
ℋ_bid.append(bid)
return action_val, tensor([action_val]), tensor([log_prob])
```

Algorithm 2: UpdatePolicy
Input: reward ∈ ℝ, P_lower, U ∈ ℝ, opp_bids: List[ℝ]
Output: None (updates π in-place)

```
if |opp_bids| < 8 then
pass
end if
ma_est ← Mean(opp_bids[-8:])
idx ← SearchSorted(𝒜, ma_est, side='left')
if idx > 0 then
target_idx ← idx - 1
else
target_idx ← 0  // Boundary case: bid minimum
end if
N ← |π|
if N > 1
denom ← lr / (N - 1)
π ← π - denom
π[target_idx] ← π[target_idx] + lr + denom
end if
π ← Clip(π, 0, 1)
π ← π / Sum(π)
```

Algorithm 3: ComputeReward
Input: my_bid, my_br, U, P_lower ∈ ℝ; opp_bids: List[ℝ]; won, tie ∈ {True, False}
Output: reward ∈ [0, 1]

```
if won then
margin ← (my_bid - P_lower) / (U - P_lower + ε)
reward ← Clip(margin, 0, 1)
else
reward ← my_br
end if
if tie then
reward ← reward × 0.5
end if
```

10. return reward

Main loop:

1. for t = 1 to T do
2. // Step 1: Observe state and opponent history
3. s_t ← GetState(my_br, opp_br, P_lower, U)
4. // Step 2: Select and submit bid
5. (action_val, action_t, log_prob_t) ← Act(s_t, training=True)
6. bid_t ← Denormalize(action_val, P_lower_init, U_init)
7. SubmitBid(bid_t)
8. // Step 3: Observe outcome
9. (won_t, opponent_bid_t, tie_t) ← ObserveTenderResult()
10. // Step 4: Compute reward
11. reward_t ← ComputeReward(bid_t, my_br, U, $\mathscr{H}$_opp, won_t, P_lower, tie_t)
12. // Step 5: Update statistics
13. total_games ← total_games + 1
14. if won_t then
15. wins ← wins + 1
16. win_rate ← wins / total_games
17. // Step 6: Record opponent behavior
18. $\mathscr{H}$_opp.append(opponent_bid_t)
19. // Step 7: Policy update (if sufficient history)
20. if |$\mathscr{H}$_bid| ≥ K then
21. UpdatePolicy(reward_t, P_lower, U, $\mathscr{H}$_opp)
22. end if
23. // Step 8: Optional logging
24. Log(t, bid_t, reward_t, win_rate, π)
25. end for
26. return π, {wins, total_games, win_rate, $\mathscr{H}$_bid, $\mathscr{H}$_opp}

4.3 Human-Centric AI Modeling for Fair and Cooperative Micro-transit Ride Prioritization

Human-Centric AI Modeling for Fair and Cooperative Microtransit Ride Prioritization

Divya Sundaresan[1][0009-0005-2680-0190], Danushka Edirimanna[2][0000-0002-5652-161X], Liwen Du[1][], Eleni Bardaka[1][0000-0001-8306-4939], Samitha Samaranayake[2][0000-0002-5459-3898], and Munindar P. Singh[1][0000-0003-3599-3893]

[1] North Carolina State University, Raleigh, NC 27606, USA
[2] Cornell University, Ithaca, NY 14850, USA

Abstract. This paper presents ongoing work on a human-centric approach to rural microtransit that combines AI-based rider modeling with system optimization to improve service efficiency without increasing rider costs. Our approach centers on a Rider Agent that estimates the likelihood that a rider will accept alternative pickup times and combines these estimates with system-level benefit to rank candidate scheduling options for flexible trips. We discuss early simulation results and we outline next steps, including LLM-supported development of alternative message framings, expanded simulation experiments, and preparation for a pilot deployment in Wilson, North Carolina.

Keywords: Civic services · Sociotechnical systems · Public ride sharing

1 Introduction

In many small communities, residents rely on microtransit to reach work, healthcare, and other essential destinations [1]. Microtransit is a shared, technology-enabled public transit service with flexible routing and pickup and dropoff locations that accommodates on-demand trip requests. Riders typically book through an app, and unlike commercial ride-hailing services such as Uber and Lyft, trips are shared and offered at a nominal fare. This model is especially relevant in suburban and rural areas, where fixed route bus systems are often costly and underused. However, public agencies still bear substantial operating costs, including payments to technology providers and the expense of vehicles and drivers, while rider fares cover only a small share of service costs.

This paper presents our vision for a smart public microtransit system that uses community-supported solutions to distribute travel demand more equitably [2]. We focus on Wilson, NC, where microtransit is the only public transit option

and many residents depend on it for daily mobility. In a recent survey, 47% of respondents reported using microtransit mainly for work trips, 86% were carless, and 57% had annual incomes below $25K [3]. Wilson's system currently receives about 20,000 trip requests per month, yet during peak hours there are long waits, delays, and missed trips for riders who often do not have any alternatives.

Our approach combines AI with schedule optimization to better satisfy rider needs under limited resources. Rather than treating rider requests as fully fixed, we model microtransit as a sociotechnical system (STS) [4], where riders, drivers, and the transit authority form the social tier, and vehicles and ride-request technologies form the technical tier. We posit that some riders may be willing to adjust pickup or dropoff times when doing so improves system performance or helps other riders. This willingness to adapt for the benefit of others is an instance of *prosociality*.

Our research question is: How can a human-centric AI system integrated with microtransit optimization encourage riders to make prosocial schedule adjustments, and to what extent can this improve system efficiency, reduce wait times, and maintain or enhance rider satisfaction without relying on pricing? We use AI to understand rider preferences and encourage such behavior in ways that can improve service for the broader community [5]. Trip requests are evaluated in real time. Inflexible trips, such as travel to work, medical appointments, or returning home, are scheduled as early as possible. Flexible trips, such as shopping, may be shifted to alternative time slots that better match rider preferences and system needs. These alternatives are identified using destination hours, learned rider preferences, and expected system benefit based on current and historical demand.

2 Rider Agent

The Rider Agent estimates the probability that a rider will accept proposed alternative pickup time slots. These estimates are combined with an estimate of benefit to the system when ranking candidate slots.

The model represents each rider as a weighted mixture over a fixed set of rider categories (e.g., student, full-time worker, part-time worker, retired). Each category c_i specifies an acceptance probability $\pi_i(s) \in [0, 1]$ for each time slot s. (We use 30-minute slots between 5 AM and 11 PM, the daily service window.)

Category-level acceptance probabilities are initialized using manually specified priors based on expected daily availability patterns. These values are shared across riders and updated gradually using aggregated data. For rider r, the predicted acceptance of slot s is computed as

$$\hat{\pi}_r(s) = \frac{\sum_i w_{ri}\, \pi_i(s)}{\sum_i w_{ri}},$$

where w_{ri} is the rider-specific weight for category c_i.

Initialization At signup, rider weights are initialized using demographic attributes such as age and employment status. These attributes determine the rider's initial distribution over the categories.

Request-Time Conditioning In addition to the rider-specific category weights, acceptance predictions are conditioned on the rider's requested pickup time t_{req}. When evaluating a candidate slot s, the model incorporates the temporal distance between s and t_{req} so that slots farther from the requested time receive lower predicted acceptance.

Slot Ranking For each candidate slot s, we compute a joint score

$$\text{score}(s) = \hat{\pi}_r(s) \cdot B(s),$$

where $B(s)$ denotes the system-level benefit of scheduling the trip in slot s. Candidate slots are ranked by this score prior to feasibility filtering and presentation to the rider.

Candidate Slot Generation For a flexible request that cannot be served immediately, we first construct a feasible time window and enumerate candidate slots within that window. For each slot s, we compute a benefit value $B(s)$ based on current and historical demand.

Depending on the policy, candidate slots are ranked either by (i) benefit alone (benefit baseline) or (ii) the joint score $\hat{\pi}_r(s) \cdot B(s)$ (learned policy). The top-ranked slots are then passed to a feasibility check, and up to three feasible alternatives.

Online Updates via Hedge After each interaction, we update rider-category weights using an exponential-weight (Hedge) rule. Let s^* be the slot selected by the rider (or no selection if all flexible options are rejected). Each category incurs a loss

$$\ell_{ri} = \begin{cases} 1 - \pi_i(s^*) & \text{if a slot is selected,} \\ \sum_{s \in A} \pi_i(s) & \text{otherwise,} \end{cases}$$

where A is the set of offered slots. Weights are updated as

$$w_{ri} \leftarrow w_{ri}\, e^{-\eta \ell_{ri}},$$

followed by normalization over valid categories.

Category Updates We use responsibility-weighted aggregation to update category acceptance probabilities. When a rider selects slot s^*, category c_i receives fractional responsibility

$$\gamma_{ri} = \frac{w_{ri}\, \pi_i(s^*)}{\sum_j w_{rj}\, \pi_j(s^*)}.$$

Each category accumulates effective counts and updates its slot acceptance probability after the effective sample size for the category exceeds a minimum threshold. This enables category models to adapt using population behavior while rider weights capture rider-specific behavior.

2.1 Simulation Assumptions

We make the following assumptions in the simulation experiments.

- **Demand input.** Trip requests and the initial start-of-day vehicle manifests are provided as an input demand manifest.
- **Rider population size and request assignment.** The rider population size is fixed (four riders). Incoming requests are assigned to riders in a round-robin manner.
- **Rider demographics.** Each rider is assigned an age (uniformly sampled from 16–75) and an employment type sampled from a fixed set: {student, working-fulltime, working-parttime, unemployed, retired}.
- **Latent time-slot preferences for riders.** Each rider has a latent preference value for each time slot, initialized from an employment-type-specific prior pattern with additional random perturbation.
- **Flexible versus inflexible trips.** Each request is labeled as flexible with probability 0.9 and inflexible with probability 0.1, independent of destination type and rider history.
- **Rider choice model under flexible shifts.** When offered alternative time slots, a rider selects one based on latent slot preferences with additional decision noise.
- **Global category models.** Category-level time-slot acceptance profiles are treated as fixed; learning updates modify only rider-specific category weights.

For the simulation, we use two kinds of learning agents and four rider agents. Rider learning agents learn the scheduling time slots and weights, initialized as in Table 1. The weights will be learned over simulation. The Category agent maintains each rider's category. A learning agent updates weights and categories. The rider agents process schedule preference probabilities and option selections. The main service agent Carma receives messages, produces options, and updates schedules for the riders.

2.2 Preliminary Results

Each simulation runs 200 requests. For baseline, we use the system-level benefit of scheduling the trip in slot s that is denoted as $B(s)$ in the agent description. The learning agent learns from every request and updates the weights. In the diagrams, these are shown as *learned*.

The agent recommends ride options that are booked and tends to propose options with longer acceptable waits and shorter detours. Figures 1a and 1b

Age	Status	Student	Full-Time	Part-Time	Unemployed	Retired
< 18	Student	0.95	0.0	0.05	0.0	0.0
< 18	Part-time	0.2	0.0	0.8	0.0	0.0
< 18	Full-time	0.2	0.8	0	0	0.0
18–64	Student	0.8	0.0	0.15	0.05	0.0
18–64	Full-time	0.0	0.8	0.15	0.05	0.0
18–64	Part-time	0.0	0.1	0.8	0.1	0.0
18–64	Unemployed	0.0	0.1	0.1	0.8	0.0
18–64	Retired	0.0	0.0	0.2	0.0	0.8
65+	Full-time	0.0	0.7	0.05	0.05	0.2
65+	Part-time	0.0	0.0	0.8	0.0	0.2
65+	Retired	0.0	0.0	0.1	0.0	0.9

Table 1: Rider-Level Priors (Full-time and part-time refer to working adults).

Demand Profile	Experiment	Service Rate	VMT	PMT	VMT/PMT	Mean Wait	Mean Detour
200 Requests	Baseline	99.5	105938.0	75226.8	1.41	833.78	533.89
	New Agent	100.0	104929	75685.9	1.39	867.95	540.50
400 Requests	Baseline	67.86	94832.1	68858.2	1.38	514.49	1199.71
	New Agent	66.07	98651.8	67082.2	1.47	474.80	2306.68

Table 2: Comparison of service metrics when optimizing with and without swap heuristic, setting the number of prebooked requests to 80 and the number of requests to be scheduled to 200.

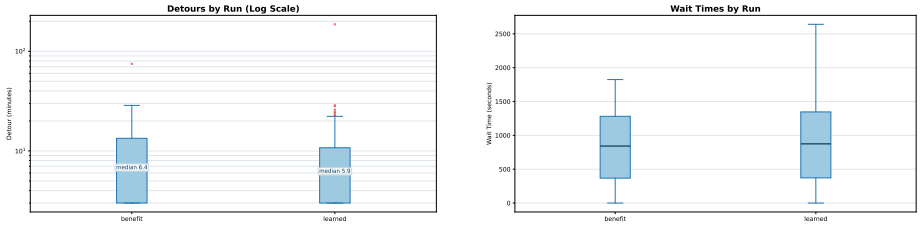
provide the detailed comparisons. However, the limited data introduces bias, resulting in a wide spread between the minimum and maximum times.

Although the flexible cases are similar across the experiments, the agent increases the average best-option value in flexible cases from 0.62 to 0.69. Because the total number of requests is small, the agent cannot meaningfully learnfrom rider information to update the category matrix; while weights do change, the differences from the baseline are minimal.

We note that these results are preliminary, and at this stage, the evaluation is intended primarily to illustrate the feasibility of the proposed framework and to provide an initial indication of how rider flexibility may be incorporated into scheduling decisions. Additional simulation experiments with larger trip request datasets are needed to assess the robustness of the approach across a broader range of demand levels, rider behavior assumptions, and operational settings.

3 Future Work

Future work will focus on improving the human-centered design of the system, strengthening the integration between AI-supported rider interactions and op-



(a) Detours (b) Wait Times

Fig. 1: "benefit" is our baseline. "learned" is our agent's performance.

timization scheduling, and advancing the proposed system from concept and prototype development toward real-world deployment. Our first priority is to conduct user studies to further refine the persuasive messaging component so that the system can more effectively encourage riders to make prosocial schedule adjustments without creating burden or reducing trust. The user studies will also allow us to test how quickly and accurately the Rider Agent can learn rider preferences and identify acceptable forms of flexibility.

As part of the user studies, we will also examine how large language models (LLMs) can support the development of alternative persuasive messages for encouraging rider flexibility. Specifically, we will use LLMs to generate message variants that differ in framing, tone, and emphasis, such as highlighting community benefit, fairness, reciprocity, or system efficiency, while keeping the requested schedule adjustment constant. We will then assess their effectiveness by comparing outcomes such as riders' willingness to accept alternative trip times, perceived fairness of the request, trust in the system, and overall satisfaction with the interaction. This process will help us identify which forms of AI-generated persuasive communication are most effective and acceptable, and how LLM-based message generation can be incorporated into a human-centered microtransit decision support framework.

Additionally, we will expand our simulation framework by testing the system under a wider range of demand scenarios and by incorporating trip request data from multiple microtransit services. This will allow us to evaluate the robustness and transferability of the proposed approach across different operating contexts. On the implementation side, future work includes completing the user interface and back-end infrastructure for our microtransit application so that the full system can support end-to-end trip requests. Together, these efforts will prepare the project for a pilot implementation in Spring 2027 in the City of Wilson, where the proposed framework can be evaluated in practice.

Acknowledgments. We acknowledge support from the US National Science Foundation (grant SCC-2325720).

References

1. Bardaka, E., Hajibabai, L., Singh, M.P.: Reimagining ride sharing: Efficient, equitable, sustainable public microtransit. IEEE Internet Computing (IC) **24**(5), 38–44 (Sep 2020). https://doi.org/10.1109/MIC.2020.3018038
2. Bardaka, E., Hentenryck, P.V., Lee, C.C., Mayhorn, C.B., Monast, K., Samaranayake, S., Singh, M.P.: Empathy and AI: Achieving equitable microtransit for underserved communities. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Special Track on *AI for Good*. pp. 7179–7187. IJCAI, Jeju, Korea (Aug 2024). https://doi.org/10.24963/ijcai.2024/794
3. Bardaka, E., Hentenryck, P.V., Lee, C.C., Mayhorn, C.B., Monast, K., Samaranayake, S., Singh, M.P.: Empathy and AI: Achieving equitable microtransit for underserved communities. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Special Track on *AI for Good*. pp. 7179–7187. IJCAI, Jeju, Korea (Aug 2024). https://doi.org/10.24963/ijcai.2024/794
4. Singh, M.P.: Norms as a basis for governing sociotechnical systems. ACM Transactions on Intelligent Systems and Technology **5**(1), 1–23 (Jan 2014). https://doi.org/10.1145/2542182.2542203, https://doi.org/10.1145/2542182.2542203
5. Sundaresan, D., Watson, A., Bardaka, E., Lee, C.C., Mayhorn, C.B., Singh, M.P.: Prosociality in microtransit. Journal of Artificial Intelligence Research (JAIR) **82**, 77–110 (Jan 2025). https://doi.org/10.1613/jair.1.16777

5 Generative and Algorithmic Influence in Sociotechnical Systems

5.1 Algorithmic Recommendations Alter the Spread of Competing Information

Algorithmic Recommendations Alter the Spread of Competing Information

João N.B. Fonseca[1[0000−0002−3804−7002]], Giuseppe M. Ferro[2,3[0000−0002−2088−4259]], and Fernando P. Santos[1[0000−0002−2310−6444]]

[1] Informatics Institute, University of Amsterdam, The Netherlands
{j.n.bastosfonseca, f.p.santos}@uva.nl
[2] Department of Ecology and Evolutionary Biology, Princeton University, USA
[3] Center for Statistics & Machine Learning, Princeton University, USA
gf6172@princeton.edu

Abstract Social media platforms are key environments for information spreading. Such information can present competing perspectives, and its reach is a matter of human action and algorithmic curation. To alleviate information overload, feed recommendation algorithms filter information based on users' interests and the perceived utility of information and engagement. Prior work studied the role of these algorithms on social dynamics such as polarisation; however, no study, to our knowledge, has focused on how this technology might alter the very dynamics of how competing information spreads. In this paper, we show preliminary results of this investigation. Through a complex systems approach, we find that algorithmic recommendations can alter the size and structural virality of information cascades, and exacerbate the reach of already attractive content. Different ways of determining what is worth recommending also seems to play a very significant role.

Keywords: social media · recommendation algorithms · complex systems · sociotechnical systems · agent-based modelling.

1 Introduction

Feed recommendation algorithms are a widespread technology in social media platforms (SMPs) [5, 20]. Given the large amount of data circulating on these platforms, pieces of information can be seen as competing with each other for users' limited attention [16, 22, 26]. Feed recommendation algorithms are a way of automatically providing each individual relevant content and of keeping them engaged for longer periods of time [4,5,17,20,21,23]. Considering this, SMPs can be seen as sociotechnical systems where users with different (and often competing) preferences have their digital environment shaped by large-scale AI tools.

While recommendations can contribute to user satisfaction due to the relevance and discovery aspects it addresses [14, 20, 21], there is an ongoing debate as to whether these are associated with a number of undesirable outcomes, such as unbalanced amplification of certain political views [9, 18] or higher exposure

to extremist content [7, 8, 27]. A number of works have previously addressed the question of *how* information spreads on SMPs [6, 10, 25]. In 2016, Goel et al. [6] introduced *structural virality* to measure how diffusions differ in terms of their structure, giving us insights about *how* information spreads. In particular, structural virality measures how much a piece of information spreads through large broadcasts or peer-to-peer transmission – if it happened mostly through the first, then structural virality is low; if it happened mostly through the latter, then structural virality is high. However, to our knowledge, the way in which feed recommendation algorithms themselves affect the scale and structure of information diffusion events is still a largely unanswered question.

With this paper, we propose answering the following two research questions: **1) do recommendations substantially alter how information spreads on SMPs?, and 2) how might this change depending on how recommendations are made?** To answer these questions, we propose a complex systems approach: by creating a simple model of the various interacting parts in these systems, we hope to gain valuable insights about the complicated algorithms used in most large SMPs. In the following sections, we will explain our approach in greater detail and present some preliminary results.

2 Methods

We assume a networked model, as networks are a common way in which information spreads on SMPs [20]. We will be using scale-free networks, a network topology that has been found to emerge on X (formerly Twitter) [15, 19]. For now, we have used networks generated by the Barabási-Albert model [2], with $N = 1.000.000$ nodes and $m = 10$, and plan on using different topologies in the future, as well as real social media networks.

There are two main components to our model. The first component is the information spreading component. For this, we will rely on the SIR epidemiological model [13], following previous work [6, 10]. In this model, one can be either Susceptible (S), Infected (I), or Recovered (R). Following [6, 12], we follow the Reed-Frost of this model – at every timestep t, every S neighbour of an I individual can be independently infected with probability β, turning into an I if the attempt is successful. By the end of that timestep, those that were infected at the start of the timestep become recovered (R). To capture information competition, the SIR needs to be extended – we adopt the simplest case, with two competing strains and perfect cross-immunity (i.e. a person infected by one strain cannot become infected by the other), as studied in [12]. The two strains, which we will call the blue and red strains, have two different infection probabilities: β_b and β_r, respectively. One can think of each strain as different posts about a topic.

Finally, a useful measure to keep track of in an SIR-like model is the *basic reproductive number*, or R_0. This represents the average number of infections an infected individual in a fully susceptible population might cause. As we will be using networks, R_0 is computed as $R_0 = \beta\langle k\rangle$, where $\langle k\rangle$ is the average degree of the network (this formula overestimates R_0 for networks with high

degree heterogeneity; while previously used by Goel et al. [6] and in network epidemiology [1], we shall use a more precise formula in the future).

The second component are the recommendations. We have simplified the complicated mechanisms of recommendation systems to the following interaction: every timestep, the recommendation algorithm has a mediating role on how likely an I individual is to contact with an S – only if contact happens, does that susceptible individual have a chance to become infected. This mechanism can be summarised with the following equation:

$$r_p(t) = \frac{M_p(t)^\alpha}{\sum_{p' \in \mathcal{P}} M_{p'}(t)^\alpha}, \tag{1}$$

where p is a post, $\mathcal{P}$ is the set of all posts, $M_p(t)$ is a certain metric of post p at timestep t, and α represents what we called the *strength of recommendation*. The value of $r_p(t)$ then represents the probability that a susceptible user will be in contact with post p through a neighbour infected with post p, at timestep t. The metric considered is popularity, measured through content diffusion on the network. Specifically, we have considered three different types of recommendations: infected-based (I-rec, where $M_p(t) = I_p(t)$), recovered-based (R-rec, where $M_p(t) = R_p(t) + 1$), and infected-recovered-based (IR-rec, where $M_p(t) = I_p(t) + R_p(t)$). I-rec can be thought of as considering the people engaged with the post p at timestep t. R-rec considers all past engagement for that post, disregarding the *current* state of engagement (this metric has to start at 1 because there is no past engagement at time step $t = 1$). Finally, IR-rec takes both past and current engagement into account.

We will study our model using simulations. Every simulation starts with a single infected node for each post – this allows us to keep track of the diffusion of the post throughout the simulation, allowing us to build its diffusion tree as the post is transmitted from node to node [6]. The initial node is either selected randomly from the whole network (*random seeding*) or randomly selected to be one of the two highest degree nodes on the network (*hub seeding*). The pseudocode we used to perform these simulations can be found in Appendix B.

3 Results

In this section, we present preliminary results from our investigation. To analyse how recommendations affect information flow on online social networks, we will be using size and structural virality (SV) as metrics [6,10,25]. Size is measured as the total number of nodes in a diffusion tree, while structural virality is measured by the average shortest path length between all pairs of nodes in a diffusion tree (see Appendix A). Finally, for each set of parameters, we have performed 50,000 simulations. Following [6], we will only consider diffusions with more than 100 infections. We will present the results for hub seeding, as there is more direct competition between the two posts, but one can see the results for random seeding in Appendix C.

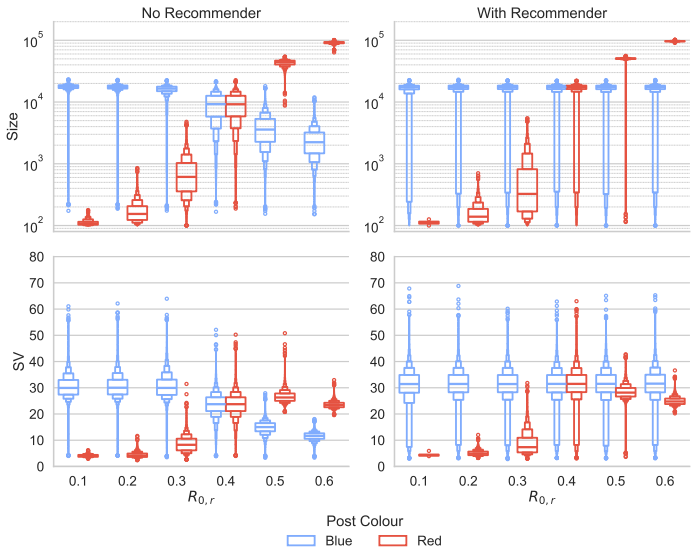


Figure 1. Letter-value plots of size and SV while varying $R_{0,r}$, in the absence or presence of an I-rec, using hub seeding ($R_{0,b} = 0.4,\ \alpha = 1$). The recommender enables blue to compete with red for higher values of $R_{0,r}$. SV is non-monotonic in both cases.

First, we will present how size and structural virality change as we change the attractiveness of the red post ($R_{0,r}$), both in presence and absence of a recommender system, while keeping the that of blue fixed at $R_{0,b} = 0.4$ (largest range of values for size and SV, using random seeding). We have assumed an I-rec with $\alpha = 1.0$ when recommendations are present. In Figure 1 one can see that when the recommender is absent, as the attractiveness of the red post increases, blue seems to have fewer and fewer chances of reaching the size it would reach if competing with a less infectious post. With I-rec, however, blue post is able to keep its size, with red still being able to grow regardless. Looking at SV, one can observe that the same pattern occurs. Moreover, SV seems to peak when $R_{0,r} = 0.4$, before decreasing again.

Our second set of results pertains to the impact of the strength of recommendation parameter (α) and the type of recommender (I- and R-rec; one can find the results for IR-rec in Appendix D). $|\alpha|$ controls how deterministic the recommendations are, and its sign determines whether to favour posts with lower or higher metrics. By observing Figure 2 and Figure 3, a common thread seems to be that the recommender tends to favour the post with highest attractiveness, even for negative values of α. As soon as α becomes positive, while both posts have a sharp increase in size, the increase of the most attractive post is far larger than that of the less attractive post. $|\alpha|$, in turn, does not seem to have a very significant effect. Furthermore, the recommenders seem to impact the spread differently: while I-rec starkly alters the diffusion as α and $R_{0,r}$ increase, R-rec

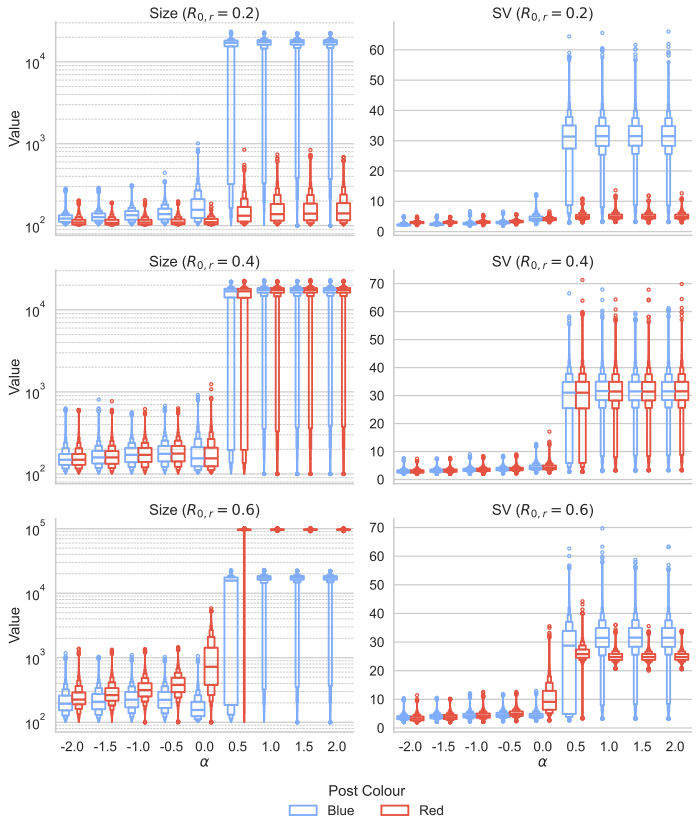


Figure 2. Letter-value plots of size and SV while varying α, with an I-rec, using hub seeding, and for three regimes of $R_{0,r}$ ($R_{0,b} = 0.4$). The recommender tends to favour the post with the highest infectivity.

affects the spread more progressively. When it comes to SV, there seem to be two different effects: I-rec seems to mostly affect SV for higher values of $R_{0,r}$, and R-rec seems to have a significant effect for the three scenarios of $R_{0,r}$.

4 Conclusion

In this paper, we show that feed recommendations can significantly affect the spread of competing information on online social networks. Although prior work has modelled competing information in various ways [16,22,26], we align our approach with studies on structural virality [6,10], modelling information spreading using a competing-strains SIR framework [12]. Recommendations seem to drastically affect not only the size of information cascades, but also their structural virality. We also find that the choice of recommender can affect information spreading in different ways.

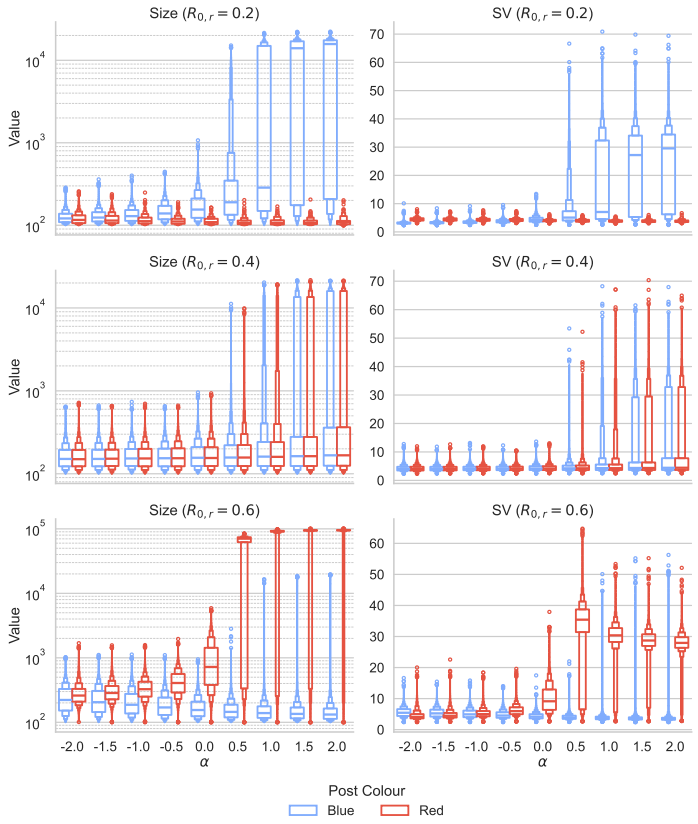


Figure 3. Letter-value plots of size and SV while varying α, with an R-rec, using hub seeding, and for three regimes of $R_{0,r}$ ($R_{0,b} = 0.4$). Higher infectivity if favoured here too, but the changes seem to be more progressive than those in Figure 2.

By using a complex systems approach, we have showed a number of preliminary results that could help one understand how feed recommendation systems might affect online information diffusion. Given the effect this technology might have on attractive information, and given that fake news usually have higher infectiousness than true news [10], it can have very serious repercussions. There are, however, potential extensions to complement our work in the future. Firstly, we would like to experiment with different network topologies, as topology can significantly alter diffusion processes (and not all online social network topologies are scale-free [3, 19, 24]). We would also like to extend the number of simulations we perform, and use more than one graph per network topology to improve the robustness of our results, as well as explore alternative types of recommenders. Finally, testing if recommendations have a significant impact on networks where there is the presence of highly homophilic minority and majority groups [11] could deliver important insights when it comes to inequalities of exposure between groups that might arise.

Acknowledgments. This research was funded by Fundação para a Ciência e Tecnologia (FCT) (grant number 2023.02219.BD).

References

1. Barabási, A.L.: Network science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**(1987) (2013)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
3. Broido, A.D., Clauset, A.: Scale-free networks are rare. Nature communications **10**(1), 1017 (2019)
4. Covington, P., Adams, J., Sargin, E.: Deep neural networks for YouTube recommendations. In: Proceedings of the 10th ACM conference on recommender systems. pp. 191–198 (2016)
5. Edelson, L., Haugen, F., McCoy, D.: Into the driver's seat with social media content feeds. Knight First Amendment Institute **25**(01) (2025)
6. Goel, S., Anderson, A., Hofman, J., Watts, D.J.: The structural virality of online diffusion. Management Science **62**(1), 180–196 (2016)
7. Haroon, M., Wojcieszak, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z.: Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. Proceedings of the National Academy of Sciences **120**(50), e2213020120 (2023)
8. Hosseinmardi, H., Ghasemian, A., Rivera-Lanas, M., Horta Ribeiro, M., West, R., Watts, D.J.: Causally estimating the effect of YouTube's recommender system using counterfactual bots. Proceedings of the National Academy of Sciences **121**(8), e2313377121 (2024)
9. Huszár, F., Ktena, S.I., O'Brien, C., Belli, L., Schlaikjer, A., Hardt, M.: Algorithmic amplification of politics on Twitter. Proceedings of the National Academy of Sciences **119**(1), e2025334119 (2022)
10. Juul, J.L., Ugander, J.: Comparing information diffusion mechanisms by matching on cascade size. Proceedings of the National Academy of Sciences **118**(46), e2100786118 (2021)
11. Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M.: Homophily influences ranking of minorities in social networks. Scientific reports **8**(1), 11077 (2018)
12. Karrer, B., Newman, M.E.: Competing epidemics on complex networks. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **84**(3), 036106 (2011)
13. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character **115**(772), 700–721 (1927)
14. Klimashevskaia, A., Jannach, D., Elahi, M., Trattner, C.: A survey on popularity bias in recommender systems. User Modeling and User-Adapted Interaction **34**(5), 1777–1834 (2024)
15. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. pp. 591–600 (2010)
16. Lerman, K.: Information is not a virus, and other consequences of human cognitive limits. Future Internet **8**(2), 21 (2016)
17. Milli, S., Belli, L., Hardt, M.: From optimizing engagement to measuring value. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 714–722 (2021)

18. Mosleh, M., Allen, J., Rand, D.G.: Divergent patterns of engagement with partisan and low-quality news across seven social media platforms. Proceedings of the National Academy of Sciences **122**(44), e2425739122 (2025)
19. Myers, S.A., Sharma, A., Gupta, P., Lin, J.: Information network or social network? the structure of the Twitter follow graph. In: Proceedings of the 23rd international conference on world wide web. pp. 493–498 (2014)
20. Narayanan, A.: Understanding social media recommendation algorithms. Knight First Amendment Institute (2023)
21. Ricci, F., Rokach, L., Shapira, B. (eds.): Recommender Systems Handbook. Springer, New York, NY, 3 edn. (Apr 2022)
22. Somazzi, A., Ferro, G.M., Garlaschelli, D., Levin, S.A.: Social media battle for attention: opinion dynamics on competing networks. arXiv preprint arXiv:2310.18309 (2023)
23. Twitter: Twitter's Recommendation Algorithm. https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm (2023), [Accessed 20-02-2026]
24. Ugander, J., Karrer, B., Backstrom, L., Marlow, C.: The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503 (2011)
25. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
26. Weng, L., Flammini, A., Vespignani, A., Menczer, F.: Competition among memes in a world with limited attention. Scientific reports **2**(1), 335 (2012)
27. Whittaker, J., Looney, S., Reed, A., Votta, F.: Recommender systems and the amplification of extremist content. Internet Policy Review **10**(2) (2021)

A Metrics

Given a diffusion tree T with N nodes, we present here the metrics we have used. Size can be defined in the following way:

$$\text{size}(T) = N, \tag{2}$$

and structural virality (SV) is defined as the average distance between all pairs of nodes in T, as defined by Goel et al. [6]:

$$\text{SV}(T) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}, \tag{3}$$

where d_{ij} is the length of the shortest path between nodes i and j.

B Simulation algorithm

Here were present the pseudocode for a single run of the simulations we performed. It follows the usual networked SIR simulation pipeline, with some additional changes to model post competition, recommendations, as well as keeping track of the diffusion trees.

Algorithm 1 Single run of a simulation

```
seed_posts(posts)      ▷ Also adds the root node to the diffusion trees of each post.

while any(active_posts(posts)) do
    for post in shuffle(posts) do
        inf_nodes ← copy(infected_nodes(post))
        rec_prob ← compute_recommendation_probability()
        for inf_node in inf_nodes do
            for sus_neigh in susceptible_neighbours(post, inf_node) do
                if rand() ≥ rec_prob then
                    continue
                else if rand() ≥ post.infection_probability then
                    continue
                end if
                infect(sus_neigh)
                add_node_to_diffusion_tree(post, inf_node, sus_neigh)
            end for
            recover_node(inf_node)
        end for
        if number_of_infected(post) = 0 then post.active ← False
        end if
    end for
end while
return get_diffusion_trees(posts)
```

C Random seeding results

In this section, we present the plots for the same type of simulations presented in Appendix 3 but using *random seeding* instead of *hub seeding*.

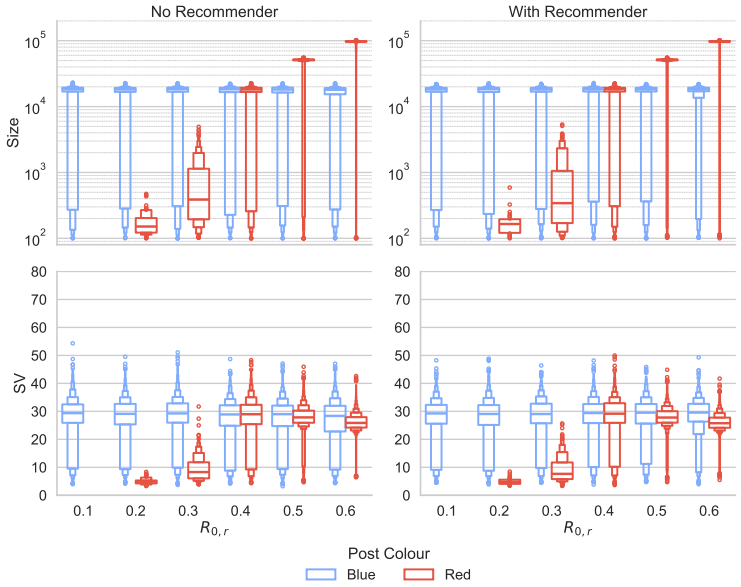


Figure 4. Letter-value plots of size and SV while varying $R_{0,r}$, in the absence or presence of an I-rec, using random seeding ($R_{0,b} = 0.4,\ \alpha = 1$). Given that in this case it is harder for diffusions to compete in the early stages, the size of the blue cascade is also largely unaffected when there is no recommender.

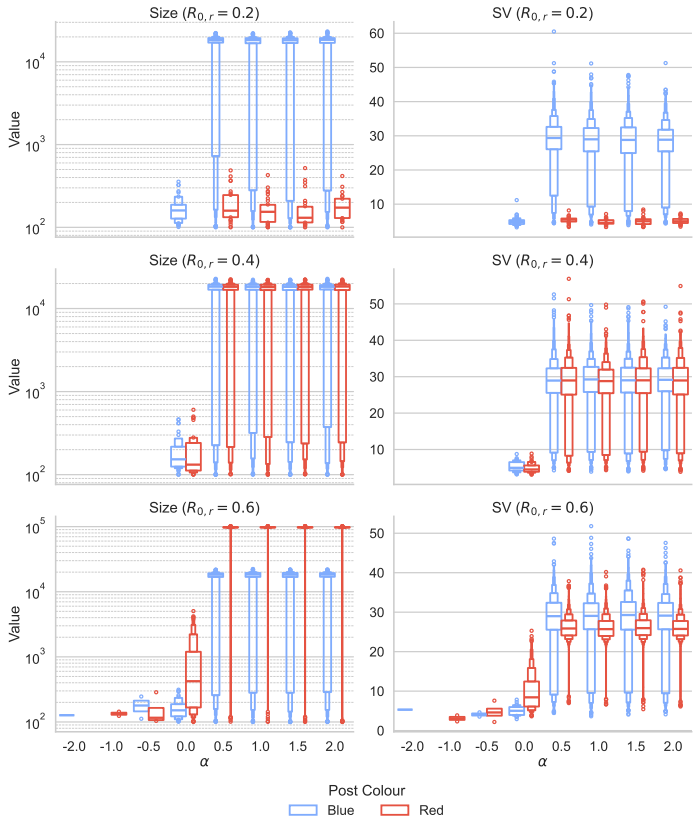


Figure 5. Letter-value plots of size and SV while varying α, with an I-rec, using random seeding, and for three regimes of $R_{0,r}$ ($R_{0,b} = 0.4$).

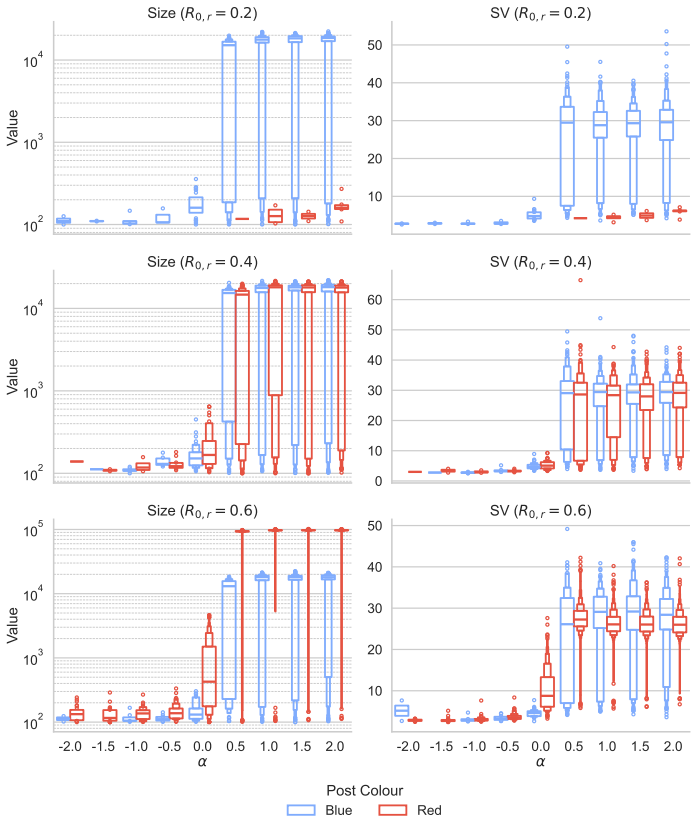


Figure 6. Letter-value plots of size and SV while varying α, with an R-rec, using random seeding, and for three regimes of $R_{0,r}$ ($R_{0,b} = 0.4$).

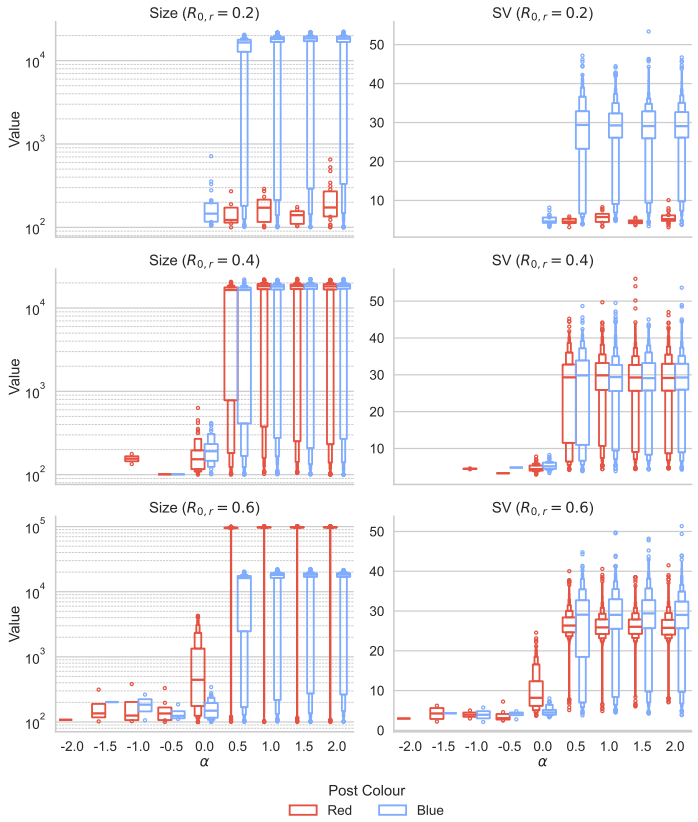


Figure 7. Letter-value plots of size and SV while varying α, with an IR-rec, using random seeding, and for three regimes of $R_{0,r}$ ($R_{0,b} = 0.4$).

D *IR*-rec results with hub seeding

In this section, we present the results for the α experiments using the IR-rec, which we did not include in the main text.

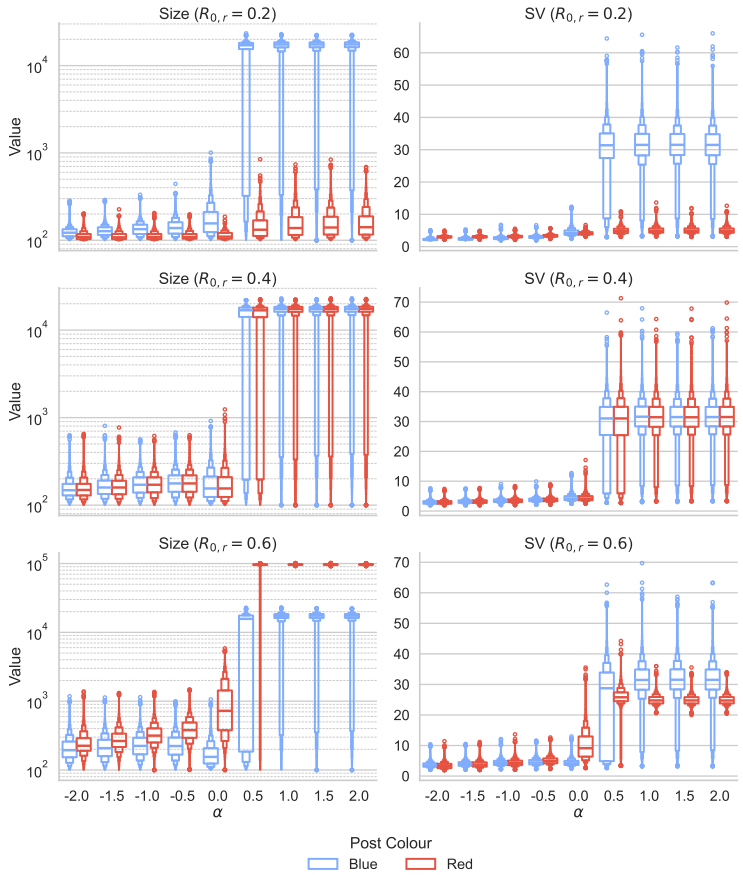


Figure 8. Letter-value plots of size and SV while varying α, with an IR-rec, using hub seeding, and for three regimes of $R_{0,r}$ ($R_{0,b} = 0.4$). The results could be seen as an in-between of those presented in the main text for I- and R-rec.

5.2 The Challenge of Realistic Personas in Generative Agent-Based Modeling

The Challenge of Realistic Personas in Generative Agent-Based Modeling

Dimitris Michailidis[1 [0000−0002−0106−1126]] and Fernando P. Santos[1 [0000−0002−2310−6444]]

Socially Intelligent Artificial Systems, University of Amsterdam
{d.michailidis, f.p.santos}@uva.nl

Abstract. Generative Agent-Based Modeling (GABM) has emerged as a paradigm shift for agent-based simulations, leveraging Large Language Models (LLMs) to create sophisticated *personas* that enrich agents with granular characteristics, emotions, and political opinions. Importantly, this new simulation paradigm has the potential to underlie sandbox systems to evaluate interventions in online social platforms. Despite the promise of these enriched simulations, we face the risk that personified agents function as caricatures rather than characters. Here we investigate this problem. We argue that highly granular personas cause agents to strictly adhere to polarizing instructions rather than engaging in human-like reasoning. To test this, we conduct an ablation study, systematically removing layers of polarizing identity from LLM-based agents. Our results indicate that agents are highly sensitive to assigned identities: minor changes in persona inputs significantly alter simulation outcomes. Specifically, when testing agents' social links formation, we find that, without stripping agents of political identification, after certain ablations, they begin to form cross-party ties and engage more with non-polarizing content. We call for more rigorous frameworks to ensure digital agents accurately reflect the nuances of human behavior.

Keywords: Generative Agent-Based Modeling · Social Simulation · Personas.

1 Introduction

The struggle for building better social media remains persistent. Despite continuous innovation in algorithms, platform architectures, and connection paradigms, researchers still struggle to address the instability of digital relationships.

Regardless of the research angle, one primary obstacle remains: evaluating interventions is incredibly hard. While the literature suggests various pathways toward pro-social digital environments [2], the friction between complex real-world social dynamics and the opaque mechanics of social media platforms makes robust, reproducible research nearly impossible.

In this context, simulations offer a valuable alternative for evaluation. By utilizing agent-based models (ABMs), researchers can construct synthetic networks where agents with specific behavioral characteristics interact in controlled

environments. These models reveal broader, macroscopic patterns by explicit user assumptions. For instance, by comparing different homophily and social influence settings [11].

The emergence of large language models (LLMs) and agentic AI offers a paradigm shift for these simulations. Generative agent-based modeling (GABM) [4], also referred to as generative social simulation, is a new subfield that has begun to leverage LLM agents to simulate networks where they consume news, share content, interact with others, and form connections [3,12,6]. This looks like a sophisticated evolution of traditional ABMs, which have long been criticized for their simplicity. Conventional agents are a collection of a limited set of attributes that take actions based on rigid rules. They lack cultural context, reasoning capabilities, and the nuances of human discourse [6]. In contrast, LLM agents can be enriched to allow for very granular heterogeneity through *personas*: mechanisms that imbue agents with physical characteristics [9,3], emotions [3], memory [9,13], professional occupations [9,3,6], political leanings [6], or personal relationships [9,14].

To ensure these personas represent real-world populations, researchers often look to empirical data, such as surveys, to extract personas from participant responses. By converting these into textual descriptions, one can theoretically imbue an LLM agent with far more granular information than traditional ABM agents. Recent studies on algorithmic feeds via these personas found that simulated networks trend toward polarization, segregation, and hierarchies [12,6,8]. These results suggest that improving social media is particularly hard [1].

Yet real-world evidence shows that pro-social behavior emerges in online social networks [2]. If humans find common ground, why do their digital equivalents fail? We ask whether this flaw lies in the persona itself, rather than the sorting algorithms. A growing body of work supports this concern: prompt-configured personas produce more stereotypical behavior than human portrayals [1], and fundamental mismatches exist between what simulations optimize for and robustness to implementation details [7]. We see a risk that these agents strictly adhere to polarizing instructions instead of performing nuanced reasoning.

To test this, we perform an ablation study. By sequentially removing layers of the agent's persona, we show that LLM agents are very sensitive to their assigned identities—even small changes to an input can significantly alter simulation outcomes. We show that, after ablations, users begin to form ties with those of the opposing party and even engage more with less polarizing content. We urge caution when interpreting results derived from personas and call for a more rigorous framework for modeling digital agents.

2 Generative Agent-Based Modeling with Personas

In generative agent-based models (GABMs), large language model (LLM) modules are integrated with individual agents, either to execute actions directly or

[1] See "Social media probably can't be fixed" `https://arstechnica.com/science/2025/08/study-social-media-probably-cant-be-fixed/`

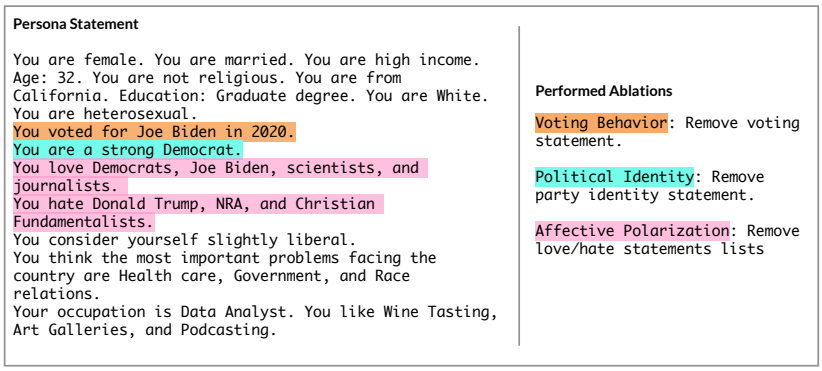


Fig. 1. To evaluate the sensitivity of generative agent-based models on personas, we ablate three layers of political identity: (1) past behavior, (2) party identification, and (3) affective polarization (love/hate lists), while keeping other statements unchanged.

to provide the memory and reasoning capabilities that support the underlying model [4]. In this paper, we employ the recently proposed generative social simulation (GSS) framework [12,10,6] to simulate a stylized, Twitter-like social network. Within this environment, agents asynchronously read and share news, interact with and repost content from others, and follow agents after an interaction. These actions are mediated by an LLM acting on behalf of each agent. By running these simulations, we can observe the longitudinal evolution of the network and quantify the extent of homophily, cross-party interaction, and polarization within the system.

Personas are the core of agents in GSS. Just as in classical ABMs agents are defined by parameters such as homophily, in GSS, they are defined by personas. For example, an agent with high homophily in ABMs becomes an agent with strong partisan identity in GSS. Similarly, a scalar opinion parameter in the range $[-1, 1]$ translates to a qualitative statement such as "You strongly believe in X." By replacing abstract variables with descriptive personas, GSS attempts, through language, to enrich an agent and capture the messy reality of human identity and interpersonal interactions.

Empirical data can be used to create realistic personas. A primary source for such data is the American National Election Studies (ANES) survey. Conducted every four years, the survey provides a rich dataset covering demographic details, political participation, party affiliation, voting behavior, and attitudes toward specific policies or public figures. By converting answers to questions into statements, an agent takes up a very specific persona. Figure 1 shows an example of such a generated persona.

3 Methodology

We use the recently proposed generative social simulation framework to create a Twitter-like platform populated by generative agents [6]. At each timestep, an agent, reads a feed of ten posts. This feed consists of five posts from accounts they follow and five selected by a "boost out partisan" algorithm, which prioritizes content from agents of the opposing political party [6]. After reading the feed, the agent chooses one of three actions: repost one of the ten visible posts, select and share a news headline from a curated database, or take no action. If an agent chooses to repost content, they are prompted to review the original author's biography and decide whether to follow them. Over time, these iterative interactions evolve the network structure of the platform. To test our hypothesis, we perform three cumulative ablations on the generated personas:

1. **Affective Polarization (AP)**: Removing phrases regarding institutions or individuals based on rated survey responses (e.g., converting a rating of $> 90/100$ from "You love X" to a neutral state).
2. **Party Identity (PID)**: Removing the statement "You are a strong/moderate Democrat/Republican/Independent.
3. **Voting Behavior (VB)**: Removing "You voted for X in 2020."

Figure 1 illustrates the content removed at each stage. These ablations are applied in two phases. First, we remove these elements only from the agent's external social bio while retaining them in the internal persona definition. This accounts for the theory that individuals self-present differently online than in person, often omitting their full range of political sentiments from their public profiles. In the second phase, we also remove these statements from the internal persona, directly influencing the agent's reasoning and actions.

Note that during the ablation, the agents are not completely rid of all leanings. They maintain political orientations (e.g., liberal or conservative) and their beliefs in societal concerns like healthcare or race relations. We only target polarizing or superficial markers, such as "love/hate" lists.

We use OpenAI's gpt-4o-mini for the LLM-driven components of our experiments. Our focus is not on evaluating the technical performance or inherent biases of specific LLMs—topics already well-documented in existing literature. Instead, we investigate the design choices involved in employing these models. We believe that our findings regarding the impact of persona design will remain relevant regardless of future improvements in LLM biases, as the core issue lies in using them to generate personas from surveys.

3.1 Experimental Setup & Evaluation

We created a platform with 500 agents and ran it for 5000 steps. Each platform setting is simulated 10 times with different seeds.

We evaluated the collective outcomes post-simulation, based on three metrics: EI index, the average clustering coefficient, and the correlation between retweets and partisanship, a measure of how being partisan benefits you in the network.

The E-I index measures the extent of homophily in the network by comparing the number of external links (EL) between agents of different parties (Democrats and Republicans) to the number of internal links (IL) within the same party.

$$EI = \frac{EL - IL}{EL + IL} \tag{1}$$

Where $EI \in [-1, 1]$, with -1 indicating perfect homophily and segregation.

We also used the clustering coefficient to measure the networks' general tendency to form tightly knit clusters, regardless of party identity. For a node v with degree k_v, the local clustering coefficient C_v is:

$$C_v = \frac{\{e_{jk} : v_j, v_k \in N(v), e_{jk} \in E\}|}{k_v(k_v - 1)}, \tag{2}$$

where $N(v)$ is the number of nodes and E are the edges. We then calculate the network-wide average clustering coefficient (average of all nodes).

Finally, to quantify the degree to which partisanship translates to social capital, we calculate the correlation between an agent's partisanship and their total received retweets. Partisanship is calculated using the ANES survey "feeling thermometer" responses [6]. The partisan score P of an agent is calculated as:

$$P = \frac{\text{feelingRepublican} - \text{feelingDemocratic}}{100} \tag{3}$$

where $P \in [-1, 1]$, with -1 indicating a strong Democrat and $+1$ a strong Republican.

4 Preliminary Results

In Figure 2, we evaluate the performance of simulated platforms across the three metrics. The top panels illustrate the ablation of persona statements from the agents' *social biographies*, while the bottom panels show the impact of removing them from the underlying *personas*.

Removing information from the publicly visible bios resulted in small, non-significant shifts. The E-I index remained negative, indicating that agents still primarily connect with co-partisans. The clustering coefficient showed no significant change across biography variations. Interestingly, removing explicit voting behavior led to a (non-significant) upward trend in partisans receiving reposts. This suggests that the internal persona dictates much of the reposting behavior of the agent, regardless of the other agents' biographies.

In contrast to the biographies, ablating the agents' internal persona instructions led to substantial changes in the network. As polarizing statements were removed, the E-I index shifted significantly. The removal of "love/hate" lists and political identity prompted a statistically significant increase in cross-party connections between Democrats and Republicans. Removing the voting behavior enhanced this trend. Furthermore, we observe an inverse trend in the clustering coefficient, where agents formed tighter, more frequent clusters. While seemingly

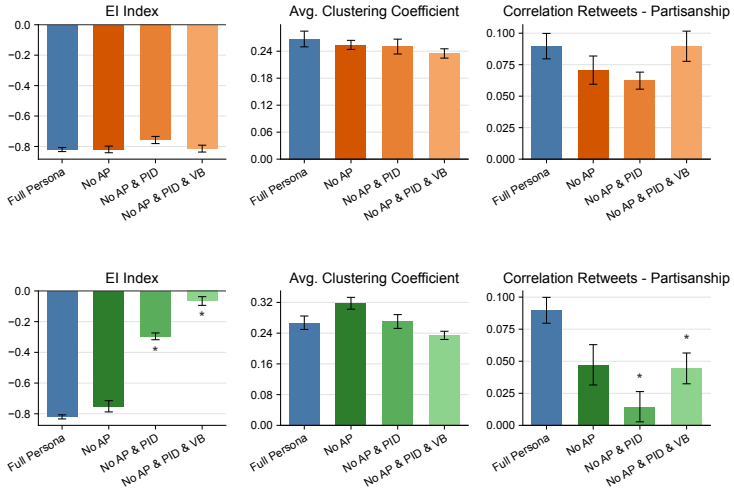


Fig. 2. We compare collective outcomes when modifying public-facing biographies (top) and internal persona prompts (bottom). Stars indicate statistically significant differences compared to the full persona baseline (Mann-Whitney test, $p < 0.05$). Results indicate that while public bio changes lead to small shifts, internal prompt ablations significantly reduce polarization, as evidenced by a rising E-I index and a shift toward penalizing extreme partisan content.

counterintuitive, this suggests that triadic closure becomes easier when agents are not strictly linked to extreme partisan personas. The internal ablation resulted in a healthier platform overall: we observed a shift toward a much smaller correlation between retweets and partisanship.

5 Conclusion

Our results call for scrutiny on using generative social simulations as human proxies. Empirical Twitter data shows a $20 - 33\%$ partisan cross-following [5], while we find that full personas produce near-zero cross-party connections. This is caused by LLMs treating persona statements as instructions to follow. We recommend mechanism ablations as standard practice and suggest enriching personas beyond survey-derived statements.

Acknowledgments

This work is supported by ERC grant (RE-LINK, 101116987, `https://doi.org/10.3030/101116987`). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Cheng, M., Durmus, E., Jurafsky, D.: Marked personas: Using natural language prompts to measure stereotypes in language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1504–1532 (2023)
2. Dörr, T., Nagpal, T., Watts, D., Bail, C.: A research agenda for encouraging prosocial behaviour on social media. Nature Human Behaviour pp. 1–9 (2025)
3. Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y.: S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984 (2023)
4. Ghaffarzadegan, N., Majumdar, A., Williams, R., Hosseinichimeh, N.: Generative agent-based modeling: an introduction and tutorial. System Dynamics Review **40**(1), e1761 (2024)
5. Halberstam, Y., Knight, B.: Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. Journal of public economics **143**, 73–88 (2016)
6. Larooij, M., Törnberg, P.: Can we fix social media? testing prosocial interventions using generative social simulation. arXiv preprint arXiv:2508.03385 (2025)
7. Li, Y., Tao, D.: Position: Ai agents are not (yet) a panacea for social simulation. arXiv preprint arXiv:2603.00113 (2026)
8. Orlando, G.M., La Gatta, V., Russo, D., Moscato, V.: Can generative agent-based modeling replicate the friendship paradox in social media simulations? In: Proceedings of the 17th ACM Web Science Conference 2025. pp. 510–515 (2025)
9. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. pp. 1–22 (2023)
10. Piao, J., Lu, Z., Gao, C., Xu, F., Hu, Q., Santos, F.P., Li, Y., Evans, J.: Emergence of human-like polarization among large language model agents. arXiv preprint arXiv:2501.05171 (2025)
11. Santos, F.P., Lelkes, Y., Levin, S.A.: Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences **118**(50), e2102141118 (2021). `https://doi.org/10.1073/pnas.2102141118`, `https://www.pnas.org/doi/abs/10.1073/pnas.2102141118`
12. Törnberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating social media using large language models to evaluate alternative news feed algorithms (2023), `https://arxiv.org/abs/2310.05984`
13. Wang, L., Zhang, J., Yang, H., Chen, Z.Y., Tang, J., Zhang, Z., Chen, X., Lin, Y., Sun, H., Song, R., et al.: User behavior simulation with large language model-based agents. ACM Transactions on Information Systems **43**(2), 1–37 (2025)
14. Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J.B., Shu, T., Gan, C.: Building cooperative embodied agents modularly with large language models. arXiv preprint arXiv:2307.02485 (2023)